Alessa Hering · Julia Schnabel · Miaomiao Zhang · Enzo Ferrante · Mattias Heinrich · Daniel Rueckert (Eds.)

LNCS 13386

Biomedical Image Registration

10th International Workshop, WBIR 2022
Munich, Germany, July 10–12, 2022
Proceedings

Springer

Lecture Notes in Computer Science 13386

More information about this series at https://link.springer.com/bookseries/558

Alessa Hering · Julia Schnabel ·
Miaomiao Zhang · Enzo Ferrante ·
Mattias Heinrich · Daniel Rueckert (Eds.)

Biomedical Image Registration

10th International Workshop, WBIR 2022
Munich, Germany, July 10–12, 2022
Proceedings

Springer

Editors
Alessa Hering
Radboud University Nijmegen Medical Center
Nijmegen, The Netherlands

Miaomiao Zhang
University of Virginia
Charlottesville, VA, USA

Mattias Heinrich
Universität zu Lübeck
Lübeck, Germany

Julia Schnabel
Helmholtz Center Munich
Neuherberg, Germany

Enzo Ferrante
CONICET, Universidad Nacional del Litoral
Santa Fe, Argentina

Daniel Rueckert
GALILEO
Technical University of Munich
Munich, Germany

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Computer Science
ISBN 978-3-031-11202-7 ISBN 978-3-031-11203-4 (eBook)
https://doi.org/10.1007/978-3-031-11203-4

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

The 10th International Workshop on Biomedical Image Registration (WBIR 2022, https://2022.wbir.info) was held in Munich, Germany, during July 10–12, 2022. After many missed conferences due to the COVID-19 pandemic, we sincerely hoped that 2022 would be a fresh restart for in-person and hybrid meetings that support exchange and collaboration within the WBIR community.

The WBIR 2022 meeting was a two-and-a-half-day workshop endorsed by the MICCAI society and supported by its new "Special Interest Group on Biomedical Image Registration" (SIG-BIR). It was organized in close spatial and temporal proximity to the Medical Imaging with Deep Learning (MIDL 2022) conference to enable interested researchers to minimize travel. Preceding editions of WBIR have run mostly as standalone two-day workshops at various locations: Bled, Slovenia (1999); Philadelphia, USA (2003); Utrecht, The Netherlands (2006); Lübeck, Germany (2010); Nashville, USA (2012); London, UK (2014); Las Vegas, USA (2016); Leiden, The Netherlands (2018), and Portorož, Slovenia (2020 / virtually). As with previous editions, the major appeal of WBIR 2022 was bringing together researchers from different backgrounds and countries, and at different points in their academic careers, who all share a great interest in image registration. Based on our relaxed two-and-a-half-day format with tutorials, three keynotes, and a main scientific program with short and long oral presentations, as well as in-person poster presentations, WBIR 2022 enabled space for lots of interaction and ample discussion among peers. As everyone's mindset is on image registration, it makes it easier for students to approach and meet their distinguished colleagues.

The WBIR 2022 proceedings, published in the Lecture Notes in Computer Science series, were established through two cycles of peer-review using OpenReview (for the first time). Full papers were reviewed in a double-blind fashion, with each submission evaluated by at least three members of the Program Committee. The Program Committee consisted of 25 experienced scientists in the field of medical image registration. All papers and reviews were afterwards discussed in an online meeting by the Paper Selection Committee to reach decisions. Short papers were categorized as either exciting early-work or abstracts of recently published/submitted long articles. Those submissions went through a lighter peer-review process, each being assigned to two members of the Paper Selection Committee. From a total of 34 submissions, 30 were selected for oral and poster presentation and 26 original works are included in these proceedings. Prominent topics include optimization, deep learning architectures, neuroimaging, diffeomorphisms, uncertainty, topology, and metrics. The presenting authors at WBIR 2022 represented a delightful diverse community with approximately 45% female speakers, nine papers from groups outside of Europe (primarily the USA), and academic levels ranging from Master's and PhD students to lecturers. To further stimulate participation from Asia, Africa, and South America we established a scholarship program that received 25 applications.

We were grateful to have three excellent keynote speakers at WBIR 2022. With rich experience in conducting numerous medical image computing projects from early feasibility to product implementation, Wolfgang Wein from ImFusion (Germany) spoke about combining visual computing with machine learning for improved registration in image-guided interventions. Maria Vakalopoulou, who is an expert on deep learning for biomedical image analysis from Paris-Saclay University (France), discussed classical and deep learning-based registration methods and their impacts on clinical diagnosis. Finally, Josien Pluim, head of the Medical Image Analysis group at Eindhoven University of Technology (The Netherlands), provided a historical overview of trends in image registration, going back to the first papers on the topic and taking us through some of the most important advances until today.

Many people contributed to the organization and success of WBIR 2022. In particular, we would like to thank the members of the Program Committee and the additional Paper Selection Committee members (Stefan Klein and Žiga Špiclin) for their work that assured the high quality of the workshop. We thank the MICCAI SIG-BIR group for their financial support and the MICCAI Society for their endorsement. Finally, we would like to thank all authors and participants of WBIR 2022 for their contributions.

June 2022

Mattias Heinrich
Alessa Hering
Julia Schnabel
Daniel Rückert
Enzo Ferrante
Miaomiao Zhang

Organization

General Chairs

Mattias Heinrich	University of Lübeck, Germany
Alessa Hering	Fraunhofer MEVIS, Germany, and Radboudumc, The Netherlands
Julia Schnabel	Helmholtz Zentrum Munich and Technical University of Munich, Germany
Daniel Rückert	Technical University of Munich, Germany

Program Committee Chairs

Enzo Ferrante	CONICET, Universidad Nacional del Litoral, Argentina
Miaomiao Zhang	University of Virginia, USA

Paper Selection Committee

Stefan Klein	Erasmus MC, The Netherlands
Žiga Špiclin	University of Ljubljana, Slovenia

Program Committee

Annkristin Lange	Fraunhofer MEVIS, Germany
Bartlomiej Papiez	University of Oxford, UK
Bernhard Kainz	Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
Tony C. W. Mok	Hong Kong University of Science and Technology, Hong Kong
Deepa Krishnaswamy	Brigham and Women's Hospital, USA
Demian Wassermann	Inria, France
Gary E. Christensen	University of Iowa, USA
Hanna Siebert	Universität zu Lübeck, Germany
Hari Om Aggrawal	Technical University of Denmark, Denmark
Ivor J. A. Simpson	University of Sussex, UK
Josien P. W. Pluim	Eindhoven University of Technology, The Netherlands
Lilla Zollei	MGH and Harvard Medical School, USA

Lucas Mansilla	CONICET, Universidad Nacional del Litoral, Argentina
Marc Niethammer	University of North Carolina at Chapel Hill, USA
Maria Vakalopoulou	CentraleSupelec, France
Mattias P. Heinrich	Universität zu Lübeck, Germany
Mirabela Rusu	Stanford University, USA
Natasa Sladoje	Uppsala University, Sweden
Pew-Thian Yap	University of North Carolina at Chapel Hill, USA
Stefan Sommer	University of Copenhagen, Denmark
Stephanie Häger	Fraunhofer MEVIS, Germany
Veronika A. Zimmer	Technische Universität München, Germany
Yi Hong	Shanghai Jiao Tong University, China

Contents

Atlases/Topology

Uncertainty

Architectures

Optimisation

Metrics/Losses

Efficiency

Atlases/Topology

Unsupervised Non-correspondence Detection in Medical Images Using an Image Registration Convolutional Neural Network

Julia Andresen[1(✉)], Timo Kepp[1], Jan Ehrhardt[1,2], Claus von der Burchard[3], Johann Roider[3], and Heinz Handels[1,2]

[1] Institute of Medical Informatics, University of Lübeck, Ratzeburger Allee 160, 23562 Lübeck, Germany
j.andresen@uni-luebeck.de
[2] German Research Center for Artificial Intelligence, Lübeck, Germany
[3] Department of Ophthalmology, Christian-Albrechts-University of Kiel, Kiel, Germany

1 Introduction

Medical image registration allows comparing images from different patients, modalities or time-points, but often suffers from missing correspondences due to pathologies and inter-patient variations. The handling of non-corresponding regions has been tackled with several approaches in the literature. For evolving processes, metamorphoses models have been used that model both spatial and appearance offsets to align images from different time-points [1,2]. Other approaches mask out [3] or weight down [4,5] the image distance measure in non-corresponding regions based on outlier detection [3], estimation of matching uniqueness [4] or correspondence probabilities [5].

Our recently published paper "Deep learning-based simultaneous registration and unsupervised non-correspondence segmentation of medical images with pathologies" [6] proposes a convolutional neural network (CNN) for joint image registration and detection of non-corresponding regions. As in previous iterative approaches [3], non-correspondences are considered as outliers in the image distance measure and are masked out. The conversion to a deep learning-based approach allows a two-step training procedure that results in better separation of spatial displacement and non-correspondence segmentation. Network training does not require manual segmentations of non-correspondences that are found in a single run, overcoming limitations of other CNN-based approaches [7–10].

A. Hering et al. (Eds.): WBIR 2022, LNCS 13386, pp. 3–7, 2022.
https://doi.org/10.1007/978-3-031-11203-4_1

2 Materials and Methods

The joint non-correspondence detection and image registration network (NCR-Net) is inspired by the U-Net [11] but follows a Y-shaped architecture with one encoder and two separate decoders. The decoders output a diffeomorphic deformation field ϕ and a non-correspondence segmentation S, respectively. Both decoders are connected to the encoder with skip connections. Moving image M and fixed image F serve as network input and outputs are generated on three resolution levels. At each resolution level, the loss function is computed to enable in-depth supervision of the network, with finer resolution levels being given more weight.

Segmentation and registration performances of NCR-Net are extensively evaluated on two datasets. The first dataset consists of longitudinal OCT images from 40 patients suffering from age-related macular degeneration. Three boundary segmentations, but no pathological labels are given for these data. The second dataset is the LPBA40 dataset, containing 40 whole-head MRI volumes from healthy probands and manual segmentations of 56 anatomical regions. To introduce known non-correspondences into the images, we simulate four different stroke lesions, two of which are quite large and the other two are smaller.

The network training takes place in two phases. First, the encoding part of the network as well as the deformation decoder are pre-trained with the "standard" objective function for image registration

$$\mathcal{L}_{\text{Reg}}(\theta; \text{M}, \text{F}) = \sum_{\mathbf{x} \in \Omega} \mathcal{D}[\text{F}, \phi \circ \text{M}] + \alpha \mathcal{R}_{\phi} + \lambda \, \mathcal{L}_{\text{opt}} \tag{1}$$

consisting of image distance measure $\mathcal{D}$ and regularization of the deformation $\mathcal{R}_{\phi}$. The last term $\mathcal{L}_{\text{opt}}$ is optional and may be used to provide supervision to the registration or segmentation task. In this work, we use the Dice loss comparing brain masks for MRI and retinal masks for OCT data in moving and fixed images to support the registration task. In the second training phase, the entire CNN is updated using

$$\mathcal{L}(\theta; \text{M}, \text{F}) = \sum_{\mathbf{x} \in \Omega} (1 - S) \cdot \mathcal{D}[\text{F}, \phi \circ \text{M}] + \alpha \mathcal{R}_{\phi} + \beta \mathcal{R}_{S} + \lambda \, \mathcal{L}_{\text{opt}} \tag{2}$$

as loss function. Here, the image distance is evaluated in corresponding regions only and the segmentation S is regularized with $\mathcal{R}_S$ consisting of segmentation volume and perimeter.

3 Results

In a first experiment, ablation studies are performed on the OCT data, comparing supervised and unsupervised versions of NCR-Net, i.e. versions trained with and without $\mathcal{L}_{\text{opt}}$, as well as versions trained with the proposed two-phase training or with loss function (2) from scratch. Two main results arise from this experiment. First, unsupervised and supervised NCR-Net perform comparably, allowing its use even for datasets without any given annotations. Second, the two-phase training scheme significantly improves Hausdorff and average surface distance of all three segmented retinal boundaries, indicating better disentanglement of spatial deformation and non-correspondence segmentation.

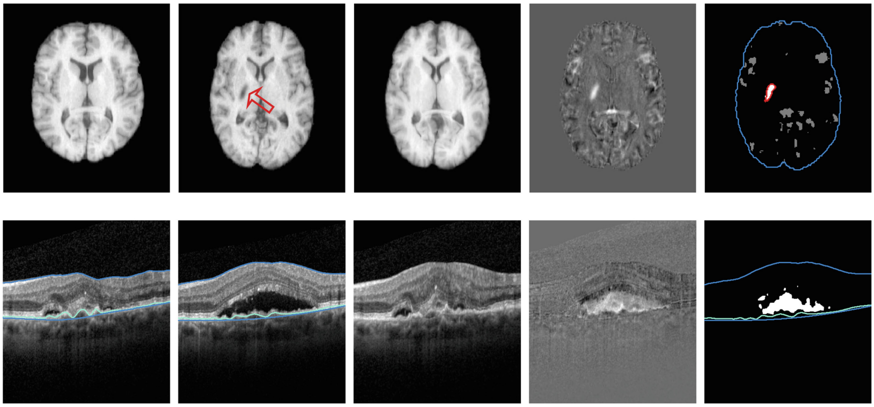


Fig. 1. Exemplary results for MRI (top row) and OCT (bottom row) data. Shown are moving, fixed, warped moving and the difference image after registration as well as the generated non-correspondence maps. Manually segmented retinal borders and automatically generated brain masks are given in blue. For the MRI data, segmentation results before and after region-growing are displayed in gray and white, respectively. The ground truth lesion is outlined in red. (Color figure online)

The registration performance of NCR-Net is further evaluated on the LPBA40 data by calculating average Jaccard indices of the given anatomical labels and comparing NCR-Net to state-of-the-art registration algorithms in 2D and 3D. NCR-Net significantly outperforms the competitive methods in the presence of large pathologies and performs comparable for images with small or no lesion. Network training with small and large simulated lesions leads to improved robustness against non-correspondences.

Finally, we evaluate the non-correspondence detection and segmentation performance of NCR-Net using the MRI data. The generated segmentations are compared to the ground truth lesion masks in two ways, first directly and second after applying region growing inside the lesions. In 2D, mean Dice scores of 0.871, 0.870, 0.630 and 0.880 are achieved for the four lesion types considered. Even though the segmentation performance in 3D is inferior, lesion detection rates are still high with 83.7 % for the worst performing lesion type.

4 Discussion

Our NCR-Net closes the gap between deep learning and iterative approaches for joint image registration and non-correspondence detection. The proposed network achieves state-of-the-art and robust registration of pathological images while additionally segmenting non-correspondent areas. With a two-step training scheme, the disentanglement of spatial deformations and non-correspondence segmentation is improved. Manual annotations may provide more supervision to the registration task, but can also be omitted without much performance loss. The simulated stroke lesions are detected as non-correspondent regions by NCR-Net very reliably and the generated segmentations are shown to be usable for unsupervised lesion segmentation and for the monitoring of evolving diseases.

References

1. Niethammer, M., et al.: Geometric metamorphosis. In: Fichtinger, G., Martel, A., Peters, T. (eds.) MICCAI 2011. LNCS, vol. 6892, pp. 639–646. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-23629-7_78
2. Rekik, I., Li, G., Wu, G., Lin, W., Shen, D.: Prediction of infant MRI appearance and anatomical structure evolution using sparse patch-based metamorphosis learning framework. In: Wu, G., Coupé, P., Zhan, Y., Munsell, B., Rueckert, D. (eds.) Patch-MI 2015. LNCS, vol. 9467, pp. 197–204. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-28194-0_24
3. Chen, K., Derksen, A., Heldmann, S., Hallmann, M., Berkels, B.: Deformable image registration with automatic non-correspondence detection. In: Aujol, J.-F., Nikolova, M., Papadakis, N. (eds.) SSVM 2015. LNCS, vol. 9087, pp. 360–371. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-18461-6_29
4. Ou, Y., Sotiras, A., Paragios, N., Davatzikos, C.: DRAMMS: deformable registration via attribute matching and mutual-saliency weighting. Med. Image Anal. **15**(4), 622–639 (2011) https://doi.org/10.1016/j.media.2010.07.002
5. Krüger, J., Schultz, S., Handels, H., Ehrhardt, J.: Registration with probabilistic correspondences-accurate and robust registration for pathological and inhomogeneous medical data. Comput. Vis. Image Underst. **190** (2020). https://doi.org/10.1016/j.cviu.2019.102839
6. Andresen, J., Kepp, T., Ehrhardt, J., von der Burchard, C., Roider, J., Handels, H.: Deep learning-based simultaneous registration and unsupervised non-correspondence segmentation of medical images with pathologies. Int. J. CARS **17**, 699–710 (2022). https://doi.org/10.1007/s11548-022-02577-4
7. Sedghi, A., Kapur, T., Luo, J., Mousavi, P., Wells, W.M.: Probabilistic image registration via deep multi-class classification: characterizing uncertainty. In: Greenspan, H., et al. (eds.) CLIP/UNSURE -2019. LNCS, vol. 11840, pp. 12–22. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32689-0_2
8. Yang, X., Kwitt, R., Styner, M., Niethammer, M.: Quicksilver: fast predictive image registration - a deep learning approach. NeuroImage **158**, 378–396 (2017)
9. Sentker, T., Madesta, F., Werner, R.: GDL-FIRE^{4D}: deep learning-based fast 4D CT image registration. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11070, pp. 765–773. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00928-1_86

10. Zhou, T., Krähenbühl, P., Aubry, M., Huang, Q., Efro, A.A.: Learning dense correspondence via 3D-guided cycle consistency. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 117–126 (2016)
11. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28

Weighted Metamorphosis for Registration of Images with Different Topologies

Anton François[1,2(✉)], Matthis Maillard[2], Catherine Oppenheim[3], Johan Pallud[3], Isabelle Bloch[2,4], Pietro Gori[2(✉)], and Joan Glaunès[1(✉)]

[1] Université de Paris-Cité, Paris, France
{anton.francois,alexis.glaunes}@parisdescartes.fr
[2] LTCI Télécom Paris, Institut Polytechnique de Paris, Paris, France
pietro.gori@parisdescartes.fr
[3] UMR 1266 INSERM, IMA-BRAIN, IPNP, Paris, France
[4] Sorbonne Université, CNRS, LIP6, Paris, France

Abstract. We present an extension of the Metamorphosis algorithm to align images with different topologies and/or appearances. We propose to restrict/limit the metamorphic intensity additions using a time-varying spatial weight function. It can be used to model prior knowledge about the topological/appearance changes (e.g., tumour/oedema). We show that our method improves the disentanglement between anatomical (i.e., shape) and topological (i.e., appearance) changes, thus improving the registration interpretability and its clinical usefulness. As clinical application, we validated our method using MR brain tumour images from the BraTS 2021 dataset. We showed that our method can better align healthy brain templates to images with brain tumours than existing state-of-the-art methods. Our PyTorch code is freely available here: https://github.com/antonfrancois/Demeter_metamorphosis.

Keywords: Image registration · Metamorphosis · Topology variation · Brain tumour

1 Introduction

When comparing medical images, for diagnosis or research purposes, physicians need accurate anatomical registrations. In practice, this is achieved by mapping images voxel wise with a plausible anatomical transformation. Possible applications are: computer assisted diagnosis or therapy, multi-modal fusion or surgical planning. These mappings are usually modelled as diffeomorphisms, as they allow for the creation of a realistic one to one deformation without modifying the topology of the source image. There exists a vast literature dealing with this subject. Some authors proposed to use stationary vectors fields, using the Lie algebra vector field exponential [1,2,14], or, more recently, Deep-Learning based methods [5,16,19,21,22,29]. Other authors used the Large Diffeomorphic

A. Hering et al. (Eds.): WBIR 2022, LNCS 13386, pp. 8–17, 2022.
https://doi.org/10.1007/978-3-031-11203-4_2

Deformation Metric Mapping (LDDMM) that uses time varying vector fields to define a right-invariant Riemannian metric on the group of diffeomorphisms. One advantage of this metric is that it can be used to build a shape space, providing useful notions of geodesics, shortest paths and distances between images [3,6,30,31]. A shortest path represents the registration between two images.

However, clinical or morphometric studies often include an alignment step between a healthy template (or atlas) and images with lesions, alterations or pathologies, like white matter multiple sclerosis or tumour. In such applications, source and target images show a different topology, thus preventing the use of diffeomorphisms, which are by definitions one-to-one mappings. Several solutions have been proposed in order to take into account such topological variations. One of the first methods was the Cost-Function Masking [7], where authors simply excluded the lesions from the image similarity cost. It is versatile and easy to implement, but it does not give good results when working with big lesions. Sdika et al. [24] proposed an inpainting method which only works on small lesions. Niethammer et al. proposed Geometric Metamorphosis [20], that combines two deformations to align pathological images which need to have the same topology. Another strategy, when working with brain images with tumours, is to use biophysical models [10,23] to mimic the growth of a tumour into an healthy image and then perform the registration (see for instance GLISTR [11]). However, this solution is slow, computationally heavy, specific to a particular kind of tumour and needs many different imaging modalities. Other works proposed to solve this problem using Deep-Learning techniques [8,12,15,25]. However, these methods strongly depend on the data-set and on the modality they have been trained on, and might not correctly disentangle shape and appearance changes.

The Metamorphic framework [13,27,30] can be seen as a relaxed version of LDDMM in which residual time-varying intensity variations are added to the diffeomorphic flow, therefore allowing for topological changes. Nevertheless, even if metamorphosis leads to very good registrations, the disentanglement between geometric and intensity changes is not unique and it highly depends on user-defined hyper-parameters. This makes interpretation of the results hard, thus hampering its clinical usage. For instance, in order to align a healthy template to an image with a tumour, one would expect that the method adds intensities only to create new structures (i.e., tumours) or to compensate for intensity changes due to the pathology (i.e. oedema). All other structures should be correctly aligned solely by the deformations. However, depending on the hyper-parameters, the algorithm might decide to account for morphological differences (i.e. mass effect of tumours) by changing the appearance rather than applying deformations. This limitation mainly comes from the fact that the additive intensity changes can theoretically be applied all over the image domain. However, in many clinical applications, one usually has prior knowledge about the position of the topological variations between an healthy image and a pathological one (e.g., tumour and oedema position).

To this end, we propose an extension of the Metamorphosis (M) model [13,27], called Weighted Metamorphosis (WM), where we introduce a time-varying spatial weight function that restricts, or limits, the intensity addition only to some specified areas. Our main contributions are: 1./ A novel time-varying spatial weight function that restricts, or limits, the metamorphic intensity additions [13,27] only to some specified areas. 2./ A new cost function that results in a set of geodesic equations similar to the ones in [13,27]. Metamorphosis can thus be seen as a specific case of our method. 3./ Evaluation on a synthetic shape dataset and on the BraTS 2021 dataset [17], proposing a simple and effective weight function (i.e., segmentation mask) when working with tumour images. 4./ An efficient PyTorch implementation of our method, available at https://github.com/antonfrancois/Demeter_metamorphosis.

2 Methods

Weighted Metamorphosis. Our model can be seen as an extension of the model introduced by Trouvé and Younès [27,30]. We will use the same notations as in [9]. Let $S, T : \Omega \rightarrow [0,1]$ be grey-scale images, where Ω is the image domain. To register S on T, we define, similarly to [27,30], the evolution of an image I_t ($t \in [0,1]$) using the action of a vector field v_t, defined as $v \cdot I_t = -\langle \nabla I_t, v_t \rangle$, and additive intensity changes, given by the residuals z_t, as:

$$\dot{I}_t = -\langle \nabla I_t, v_t \rangle + \mu M_t z_t, \qquad \text{s.t. } I_0 = S,\ I_1 = T,\ \mu \in \mathbb{R}^+. \tag{1}$$

where we introduce the weight function $M_t : x \in \Omega \rightarrow [0,1]$ (at each time $t \in [0,1]$) that multiplies the residuals z_t at each time step t and at every location x. We assume that M_t is smooth with compact support and that it can be fully computed before the optimisation. Furthermore, we also define a new pseudo-norm $\| \bullet \|_{M_t}$ for z. Since we want to consider the magnitude of z only at the voxels where the intensity is added, or in other terms, where $M_t(x)$ is not zero, we propose the following pseudo-norm:

$$\|z_t\|_{M_t}^2 = \left\| \sqrt{M_t} z_t \right\|_{L^2}^2 = \langle z_t, M_t z_t \rangle_{L^2} \tag{2}$$

This metric will sum up the square values of z inside the support of M_t. As usual in LDDMM, we assume that each $v_t \in V$, where V is a Hilbert space with a reproducing kernel K_σ, which is chosen here as a Gaussian kernel parametrized by σ [18,28]. Similarly to [27,30], we use the sum of the norm of z and the one of v (i.e., the total kinetic energy), balanced by ρ, as cost function:

$$E_{\mathrm{WM}}(v, I) = \int_0^1 \|v_t\|_V^2 + \rho \|z_t\|_{M_t}^2 dt, \quad \text{s.t. } I_0 = S,\ I_1 = T,\ \rho \in \mathbb{R}^+ \tag{3}$$

where z depends on I through Eq. 1. By minimising Eq. 3, we obtain an exact matching.

Theorem 1. *The geodesics associated to Eq. 3 are:*

$$\begin{cases} v_t = -\frac{\rho}{\mu} K_\sigma \star (z_t \nabla I_t) \\ \dot{z}_t = - \quad \nabla \cdot (z_t v_t) \\ \dot{I}_t = -\langle \nabla I_t, v_t \rangle + \mu M_t z_t \end{cases} \tag{4}$$

where $\nabla \cdot (v)$ is the divergence of the field v and $\star$ represents the convolution.

Proof. This proof is similar to the one in [30], Chap. 12, but needs to be treated carefully due to the pseudo-norm $\|z_t\|^2_{M_t} = \left\langle z_t, \frac{1}{\mu}(\dot{I}_t + v_t \cdot \nabla I_t) \right\rangle_{L^2}$. We aim at computing the variations of Eq. 3 with respect to I and v and compute the Euler-Lagrange equations. To this end, we define two Lagrangians: $L_I(t, I, \dot{I}) = E_{\mathrm{WM}}(\bullet, v)$ and $L_v(t, v, \dot{v}) = E_{\mathrm{WM}}(I, \bullet)$ and start by computing the variations h with respect to v:

$$D_v L_v \cdot h = \int_0^1 \langle 2(K^{-1} v_t + \frac{\rho}{\mu} z_t \nabla I_t), h_t \rangle_{L^2} dt \tag{5}$$

Then, noting that $\nabla_v L_v = 2(K^{-1} v_t + \frac{\rho}{\mu} z_t \nabla I_t)$ and since $\nabla_{\dot{v}} L_v = 0$, the Euler-Lagrange equation is:

$$\nabla_v L_v - \dot{\nabla}_{\dot{v}} L_v = 0 \Leftrightarrow v_t = -\frac{\rho}{\mu} K \star (z_t \nabla I_t) \tag{6}$$

as in the classical Metamorphosis framework [30]. Considering the variation of I, we have $D_I \left\| \sqrt{M_t} z \right\|^2_{L^2} = \left\langle z_t, \frac{1}{\mu} v_t \cdot \nabla h_t \right\rangle_{L^2}$, thus obtaining:

$$D_I L_I \cdot h = 2 \int_0^1 \left\langle z_t, \frac{1}{\mu} \nabla h_t \cdot v_t \right\rangle_{L^2} dt = \int_0^1 \left\langle -\frac{2}{\mu} \nabla \cdot (z_t v_t), h_t \right\rangle_{L^2} dt \tag{7}$$

and $D_{\dot{I}} L_I \cdot h = \int_0^1 \langle \frac{2}{\mu} z_t, h_t \rangle_{L^2} dt$. We deduce that $\nabla_I L_I = \frac{2}{\mu} \nabla \cdot (z_t v_t)$ and as $\nabla_{\dot{v}} L_v = \frac{2}{\mu} z_t$, its Euler-Lagrange equation is:

$$\nabla_I L_I - \dot{\nabla}_{\dot{I}} L_I = 0 \Leftrightarrow \dot{z}_t = -\nabla \cdot (z_t v_t) \tag{8}$$

We can first notice that, by following the geodesic paths, the squared norms over time are conserved ($\forall t \in [0, 1], \|v_0\|^2_V = \|v_t\|^2_V$) and thus one can actually optimise using only the initial norms. Furthermore, since v_0 can be computed from z_0 and I_0, the only parameters of the system are z_0 and I_0. As it is often the case in the image registration literature, we propose to convert Eq. 3 into an unconstrained inexact matching problem, thus minimising:

$$J_{\mathrm{WM}}(z_0) = \|I_1 - T\|^2_{L_2} + \lambda \left[\|v_0\|^2_V + \rho \|z_0\|^2_M \right], \quad \lambda \in \mathbb{R}^+, I_0 = S \tag{9}$$

where I_1 is integrated with Eq. 4, $\|v_0\|^2_V = \langle z_0 \nabla S, K_\sigma \star (z_0 \nabla S) \rangle$ and λ is the trade-off between the data term (based here on a L2-norm, but any metric could be used as well) and the total regularisation.

Weighted Function Construction. The definition of the weight function M_t is quite generic and could be used to register any kind of topological/appearance differences. Here, we restrict to brain tumour images and propose to use an evolving segmentation mask as weight function. We assume that we already have the binary segmentation mask B of the tumour (comprising both oedema and necrosis) in the pathological image and that healthy and pathological images are rigidly registered, so that B can be rigidly moved onto the healthy image. Our goal is to obtain an evolving mask $M_t : [0, 1] \times \Omega \to [0, 1]$ that somehow mimics the tumour growth in the healthy image starting from a smoothed small ball in the centre of the tumour (M_0) and smoothly expanding it towards B. We generate M_t by computing the LDDMM registration between M_0 and B. Please note that here one could use an actual biophysical model [10,23] instead of the proposed simplistic approximation based on LDDMM. However, it would require prior knowledge, correct initialisation and more than one imaging modality. The main idea is to smoothly and slowly regularise the transformation so that the algorithm first modifies the appearance only in a small portion of the image, trying to align the surrounding structure only with deformations. In this way, the algorithm tries to align all structures with shape changes adding/removing intensity only when necessary. This should prevent the algorithm from changing the appearance instead of applying deformations (i.e. better disentanglement) and avoid wrong overlapping between new structures (e.g. tumour) and healthy ones. Please refer to Fig. 1 for a visual explanation.

3 Results and Perspectives

Implementation Details. Our Python implementation is based on PyTorch for automatic differentiation and GPU support, and it uses the semi-Lagrangian formulation for geodesic shooting presented in [9]. For optimisation we use the PyTorch's Adadelta method.

Synthetic Data. Here, we illustrate our method on a 300 × 300 grey-scale image registration toy-example (Fig. 1). We can observe the differences in the geodesic image evolution for LDDMM, Metamorphosis (M) and Weighted Metamorphosis (WM) with a constant and evolving mask. First, LDDMM cannot correctly align all grey ovals and Metamorphosis results in an image very similar to the target. However, most of the differences are accounted for with intensity changes rather than deformations. By contrast, when using the proposed evolving mask (fourth row), the algorithm initially adds a small quantity of intensity in the middle of the image and then produces a deformation that enlarges it and correctly pushes away the four grey ovals. In the third row, a constant mask ($M_t = M_1, \forall t \in [0, 1]$) is applied. One can observe that, in this case, the bottom and left ovals overlap with the created central triangle and therefore pure deformations cannot correctly match both triangle and ovals. In all methods, the registration was done with the same field smoothness regularisation σ and integration steps. Please note that the four grey ovals at the border are not correctly matched with LDDMM and, to a lesser extent, also with our method. This is due to the

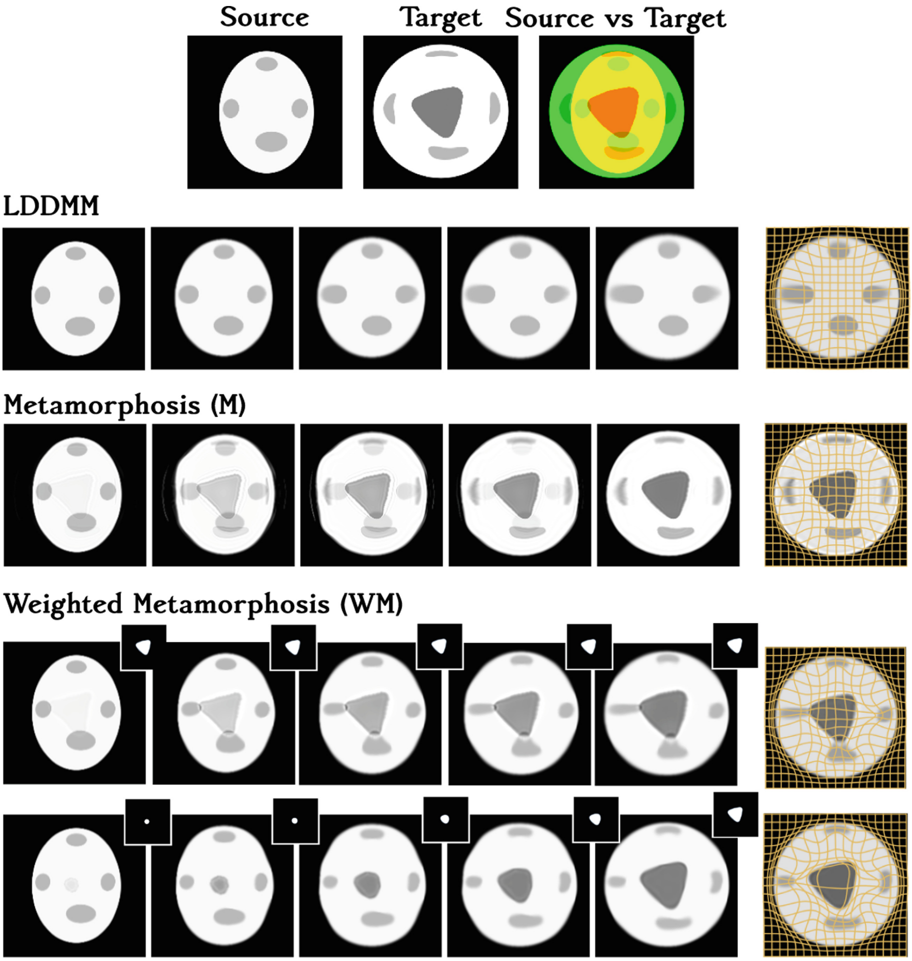


Fig. 1. Comparison between LDDMM, Metamorphosis and our method. Image registration toy example. Differently from the Source image (S), the Target image (T) has a big central triangle that has grown "pushing" the surroundings ovals. Note that the bottom and left ovals in S overlap with the triangle in T. The two last rows show our method using a constant and time evolving mask (see Sect. 2). The used mask is displayed on the top right corner of each image. `see animations in GitHub in notebook : toyExample_weightedMetamorphosis.ipynb`

L2-norm data term since these shapes do not overlap between the initial source and target images and therefore the optimiser cannot match them.

Validation on 2D Real Data. For evaluation, we used T1-w MR images from the BraTS 2021 dataset [4,17]. For each patient, a tumour segmentation is provided. We selected the same slice for 50 patients resizing them to 240 × 240 and making sure that a tumour was present. We then proceeded to register the

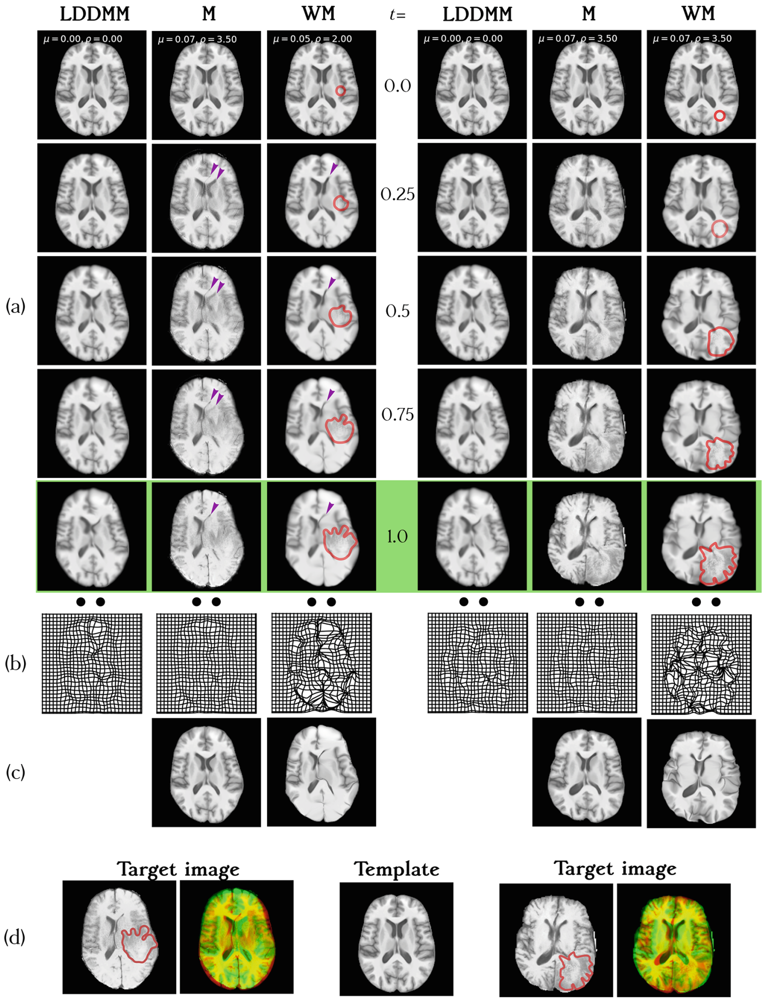


Fig. 2. Registrations on MRI brains presenting brain tumours. Two examples from BraTS database [4,17]. Comparison of geodesic shooting for LDDMM, Metamorphosis (M) and Weighted Metamorphosis (WM). (a&d) On the target images and the geodesic integration, the temporal mask is indicated by the red outline. The final result of each integration can be seen in the green outlined row. (b) The deformation grids retrieved from each method and (c) the template image deformed without intensity additions for each concerned method. Purple arrows in columns 2 and 3 in the top right part of each image show the evolution of one ventricle through registration: while M makes the ventricle disappear and reappear, WM coherently displaces the structure. (d) Target images with the segmentation outlined in red; the colored image is its superposition with the source. `see animations in GitHub in notebook : brains_weightedMetamorphosis.ipynb`

healthy brain template SRI24 [26] to each of the selected slices (see Fig. 2 for two examples). To evaluate the quality of the alignment we used three different measures in Table 1: 1./ the Sum of Squared Differences (SSD) (i.e. L2-norm) between the target (T) and the transformed source (S) images. This is a natural choice as it is used in the cost function. 2./ the SSD between T and the deformed S without considering intensity changes. This is necessary since Metamorphoses could do a perfect matching without using deformations but only intensity changes. 3./ A Dice score between the segmentations of the ventricles in the deformed S and T. The ventricles were manually segmented. All methods should correctly align the ventricles using solely pure deformations since theses regions are (theoretically) not infiltrated by the tumour (*i.e.*, no intensity modifications) and they can only be displaced by the tumor mass effect.

Table 1. Quantitative evaluation for different registration methods. Results were computed on a test set of 50 2D 240 × 240 images from BraTS 2021 dataset. - $(*)$ SSD for CFM is computed over the domain outside the mask.

Method	LDDMM [18]	Meta. [9]	WM (ours)	MAE [8]	Voxelm. [5]	CFM [7]
SSD (final)	223 ± 51	**36 ± 9**	65 ± 71	497 ± 108	166.71 ± 37	49* ± 28
SSD (def.)	–	112 ± 21	**102 ± 76**	865 ± 172	–	–
Dice score	68.6 ± 11.9	74.1 ± 9.3	**77.2 ± 10.1**	60.6 ± 8.79	66.8 ± 10	45.0 ± 13.5

We compared our method with LDDMM [6], Metamorphosis [27], using the implementation of [9], Metamorphic Auto-Encoder (MAE) [8], Voxelmorph [5] and Cost Function Masking (CFM) [7] (see Table 1). Please note that we did not include other deep-learning methods, such as [12,15], since they only work the other way around, namely they can only register images with brain tumours to healty templates. As expected, Metamorphosis got the best score for SSD (final) as it is the closest to an exact matching method. However, WM outperformed all methods in terms of Dice score obtaining a very low SSD (both final and deformation-only). This means that our method correctly aligned the ventricles, using only the deformation, and at the same time it added intensity only where needed to globally match the two images (i.e., good disentanglement between shape and appearance).

Perspectives and Conclusion. In this work, we introduced a new image registration method, Weighted Metamorphosis, and showed that it successfully disentangles deformation from intensity addition in metamorphic registration, by using prior information. Furthermore, the use of a spatial mask makes our method less sensitive to hyper-parameter choice than Metamorphosis, since it spatially constrains the intensity changes. We also showed that WM improves the accuracy of registration of MR images with brain tumours from the BRATS 2021 dataset. We are confident that this method could be applied to any kind of medical images showing exogenous tissue growth with mass-effect. A future research direction will be the integration of methods from topological data analysis, such as persistent homology, to improve even more the disentanglement between geometric and appearance changes. We also plan to adapt our method to 3D data.

Acknowledgement. M. Maillard was supported by a grant of IMT, Fondation Mines-Télécom and Institut Carnot TSN, through the "Futur & Ruptures" program.

References

1. Arsigny, V., Commowick, O., Pennec, X., Ayache, N.: A log-euclidean framework for statistics on diffeomorphisms. In: Larsen, R., Nielsen, M., Sporring, J. (eds.) MICCAI 2006. LNCS, vol. 4190, pp. 924–931. Springer, Heidelberg (2006). https://doi.org/10.1007/11866565_113
2. Ashburner, J.: A fast diffeomorphic image registration algorithm. NeuroImage **38**(1), 95–113 (2007)
3. Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C.: Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Med. Image Anal. **12**(1), 26–41 (2008)
4. Baid, U., et al.: The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv:2107.02314 (2021)
5. 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)
6. Beg, M.F., Miller, M.I., Trouvé, A., Younes, L.: Computing large deformation metric mappings via geodesic flows of diffeomorphisms. Int. J. Comput. Vis. **61**(2), 139–157 (2005)
7. Brett, M., Leff, A., Rorden, C., Ashburner, J.: Spatial normalization of brain images with focal lesions using cost function masking. NeuroImage **14**(2), 486–500 (2001)
8. Bône, A., Vernhet, P., Colliot, O., Durrleman, S.: Learning joint shape and appearance representations with metamorphic auto-encoders. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12261, pp. 202–211. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59710-8_20
9. François, A., Gori, P., Glaunès, J.: Metamorphic image registration using a semi-lagrangian scheme. In: SEE GSI (2021)
10. Gooya, A., Biros, G., Davatzikos, C.: Deformable registration of glioma images using EM algorithm and diffusion reaction modeling. IEEE Trans. Med. Imaging **30**(2), 375–390 (2011)
11. Gooya, A., Pohl, K., Bilello, M., Cirillo, L., Biros, G., Melhem, E., Davatzikos, C.: GLISTR: glioma image segmentation and registration. IEEE Trans. Med. Imaging **31**, 1941–54 (2012)
12. Han, X., et al.: A deep network for joint registration and reconstruction of images with pathologies. In: Liu, M., Yan, P., Lian, C., Cao, X. (eds.) MLMI 2020. LNCS, vol. 12436, pp. 342–352. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59861-7_35
13. Holm, D.D., Trouvé, A., Younes, L.: The euler-poincaré theory of metamorphosis. Q. Appl. Math. **67**(4), 661–685 (2009)
14. Lorenzi, M., Ayache, N., Frisoni, G.B., Pennec, X.: LCC-Demons: a robust and accurate symmetric diffeomorphic registration algorithm. NeuroImage **81**, 470–483 (2013)

15. Maillard, M., François, A., Glaunès, J., Bloch, I., Gori, P.: A deep residual learning implementation of metamorphosis. In: IEEE ISBI (2022)
16. Mansilla, L., Milone, D.H., Ferrante, E.: Learning deformable registration of medical images with anatomical constraints. Neural Netw. **124**, 269–279 (2020)
17. Menze, B.H., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE Trans. Med. Imaging **34**(10), 1993–2024 (2015)
18. Miller, M.I., Trouvé, A., Younes, L.: Geodesic shooting for computational anatomy. J. Math. Imaging Vis. **24**(2), 209–228 (2006)
19. Mok, T.C.W., Chung, A.C.S.: large deformation diffeomorphic image registration with laplacian pyramid networks. In: MICCAI (2020)
20. Niethammer, M., et al.: Geometric Metamorphosis. In: Fichtinger, G., Martel, A., Peters, T. (eds.) MICCAI 2011. LNCS, vol. 6892, pp. 639–646. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-23629-7_78
21. Niethammer, M., Kwitt, R., Vialard, F.X.: Metric learning for image registration. In: CVPR, pp. 8455–8464 (2019)
22. Rohé, M.M., Datar, M., Heimann, T., Sermesant, M., Pennec, X.: SVF-Net: learning deformable image registration using shape matching. In: MICCAI, p. 266 (2017)
23. Scheufele, K., et al.: Coupling brain-tumor biophysical models and diffeomorphic image registration. Comput. Methods Appl. Mech Eng. **347**, 533–567 (2019)
24. Sdika, M., Pelletier, D.: Nonrigid registration of multiple sclerosis brain images using lesion inpainting for morphometry or lesion mapping. Hum. Brain Mapp. **30**(4), 1060–1067 (2009)
25. Shu, Z., et al.: Deforming autoencoders: unsupervised disentangling of shape and appearance. In: ECCV (2018)
26. Torsten, R., Zahr, N.M., Sullivan, E.V., Pfefferbaum, A.: The SRI24 multichannel atlas of normal adult human brain structure. Hum. Brain Mapp. **31**(5), 798–819 (2010)
27. Trouvé, A., Younes, L.: Local geometry of deformable templates. SIAM J. Math. Anal. **37**(1), 17–59 (2005)
28. Vialard, F.X., Risser, L., Rueckert, D., Cotter, C.J.: Diffeomorphic 3D image registration via geodesic shooting using an efficient adjoint calculation. Int. J. Comput. Vis. **97**(2), 229–241 (2011)
29. Yang, X., Kwitt, R., Styner, M., Niethammer, M.: Quicksilver: fast predictive image registration - a deep learning approach. NeuroImage **158**, 378–396 (2017)
30. Younes, L.: Deformable objects and matching functionals. In: Shapes and Diffeomorphisms. AMS, vol. 171, pp. 243–289. Springer, Heidelberg (2019). https://doi.org/10.1007/978-3-662-58496-5_9
31. Zhang, M., Fletcher, P.T.: Fast diffeomorphic image registration via fourier-approximated lie algebras. Int. J. Comput. Vis. **127**(1), 61–73 (2018)

LDDMM Meets GANs: Generative Adversarial Networks for Diffeomorphic Registration

Ubaldo Ramon(✉), Monica Hernandez, and Elvira Mayordomo

Department of Computer Science and Systems Engineering, School of Engineering and Architecture, University of Zaragoza, Zaragoza, Spain
{uramon,mhg,elvira}@unizar.es

Abstract. The purpose of this work is to contribute to the state of the art of deep-learning methods for diffeomorphic registration. We propose an adversarial learning LDDMM method for pairs of 3D mono-modal images based on Generative Adversarial Networks. The method is inspired by the recent literature on deformable image registration with adversarial learning. We combine the best performing generative, discriminative, and adversarial ingredients from the state of the art within the LDDMM paradigm. We have successfully implemented two models with the stationary and the EPDiff-constrained non-stationary parameterizations of diffeomorphisms. Our unsupervised learning approach has shown competitive performance with respect to benchmark supervised learning and model-based methods.

Keywords: Large deformation diffeomorphic metric mapping · Generative Adversarial Networks · Geodesic shooting · Stationary velocity fields

1 Introduction

Since the 80s, deformable image registration has become a fundamental problem in medical image analysis [1]. A vast literature on deformable image registration methods exists, providing solutions to important clinical problems and applications. Up to the ubiquitous success of methods based on Convolutional Neural Networks (CNNs) in computer vision and medical image analysis, the great majority of deformable image registration methods were based on energy minimization models [2]. This traditional approach is model-based or optimization-based, in contrast with recent deep-learning approaches that are

U. Ramon, M. Hernandez and E. Mayordomo—With the ADNI Consortium.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-11203-4_3.

A. Hering et al. (Eds.): WBIR 2022, LNCS 13386, pp. 18–28, 2022.
https://doi.org/10.1007/978-3-031-11203-4_3

known as learning-based or data-based. Diffeomorphic registration constitutes the inception point in Computational Anatomy studies for modeling and understanding population trends and longitudinal variations, and for establishing relationships between imaging phenotypes and genotypes in Imaging Genetics [3,4]. Model-based diffeomorphic image registration is computationally costly. In fact, the huge computational complexity of large deformation diffeomorphic metric mapping (LDDMM) [5] is considered the curse of diffeomorphic registration, where very original solutions such as the stationary parameterization [6–8], the EPDiff constraint on the initial velocity field [9], or the band-limited parameterization [10] have been proposed to alleviate the problem. Since the advances that made it possible to learn the optical flow using CNNs (FlowNet [11]), dozens of deep-learning data-based methods have been proposed to approach the problem of deformable image registration in different clinical applications [12], some specifically for diffeomorphic registration [13–22]. Overall, all data-based methods yield fast inference algorithms for diffeomorphism computation once the difficulties with training have been overcome. Generative Adversarial Networks (GANs) is an interesting unsupervised approach where some interesting proposals for non-diffeomorphic deformable registration have been made [23] (2D) and [24,25] (3D). GANs have also been used for diffeomorphic deformable template generation [26], where the registration sub-network is based on an established U-net architecture [22,27], or for finding deformations for other purposes like interpretation of disease evidence [28]. A GAN combines the interaction of two different networks during training: a generative network and a discrimination network. The generative network itself can be regarded as an unsupervised method that, once included in the GAN system, is trained with the feedback of the discrimination network. The discriminator helps further update the generator during training with information regarding how the appearance of plausible warped source images. The main contribution of this work is the proposal of a GAN-based unsupervised learning LDDMM method for pairs of 3D mono-modal images, the first to use GANs for diffeomorphic registration. The method is inspired by the recent literature for deformable image registration with adversarial learning [24,25] and combines the best performing components within the LDDMM paradigm. We have successfully implemented two models for the stationary and the EPDiff-constrained non-stationary parameterizations and demonstrate the effectiveness of our models in both 2D simulated and 3D real brain MRI data.

2 Background on LDDMM

Let $\Omega \subseteq \mathbb{R}^d$ be the image domain. Let $Diff(\Omega)$ be the LDDMM Riemannian manifold of diffeomorphisms and V the tangent space at the identity element. $Diff(\Omega)$ is a Lie group, and V is the corresponding Lie algebra [5]. The Riemannian metric of $Diff(\Omega)$ is defined from the scalar product in V, $\langle v, w\rangle_V = \langle Lv, w\rangle_{L^2}$, where L is the invertible self-adjoint differential operator associated with the differential structure of $Diff(\Omega)$. In traditional LDDMM methods, $L = (Id - \alpha \varDelta)^s, \alpha > 0, s \in \mathbb{R}$ [5]. We will denote with K the inverse of operator L. Let I_0 and I_1 be the source and the target images. LDDMM is formulated from the minimization of the variational problem

$$E(v) = \frac{1}{2} \int_0^1 \langle Lv_t, v_t \rangle_{L^2} dt + \frac{1}{\sigma^2} \| I_0 \circ (\phi_1^v)^{-1} - I_1 \|_{L^2}^2. \tag{1}$$

The LDDMM variational problem was originally posed in the space of time-varying smooth flows of velocity fields, $v \in L^2([0,1], V)$. Given the smooth flow $v : [0,1] \rightarrow V$, $v_t : \varOmega \rightarrow \mathbb{R}^d$, the solution at time $t = 1$ to the evolution equation

$$\partial_t (\phi_t^v)^{-1} = -v_t \circ (\phi_t^v)^{-1} \tag{2}$$

with initial condition $(\phi_0^v)^{-1} = id$ is a diffeomorphism, $(\phi_1^v)^{-1} \in Diff(\varOmega)$. The transformation $(\phi_1^v)^{-1}$, computed from the minimum of $E(v)$, is the diffeomorphism that solves the LDDMM registration problem between I_0 and I_1. The most significant limitation of LDDMM is its large computational complexity. In order to circumvent this problem, the original LDDMM variational problem is parameterized on the space of initial velocity fields

$$E(v_0) = \frac{1}{2} \langle Lv_0, v_0 \rangle_{L^2} + \frac{1}{\sigma^2} \| I_0 \circ (\phi_1^v)^{-1} - I_1 \|_{L^2}^2. \tag{3}$$

where the time-varying flow of velocity fields v is obtained from the EPDiff equation

$$\partial_t v_t + K[(Dv_t)^T \cdot Lv_t + DLv_t \cdot v_t + Lv_t \cdot \nabla \cdot v_t] = 0 \tag{4}$$

with initial condition v_0 (geodesic shooting). The diffeomorphism $(\phi_1^v)^{-1}$, computed from the minimum of $E(v_0)$ via Eqs. 4 and 2, verifies the momentum conservation constraint (MCC) [29], and, therefore, it belongs to a geodesic path on $Diff(\varOmega)$. Simultaneously to the MCC parameterization, a family of methods was proposed to further circumvent the large computational complexity of the original LDDMM [6–8]. In all these methods, the time-varying flow of velocity fields v is restricted to be steady or stationary [30]. In this case, the solution does not belong to a geodesic.

3 Generative Adversarial Networks for LDDMM

Similarly to model-driven approaches for estimating LDDMM diffeomorphic registration, data-driven approaches for learning LDDMM diffeomorphic registration aim at the inference of a diffeomorphism $(\phi_1^v)^{-1}$ such that the LDDMM energy is minimized for a given (I_0, I_1) pair. In particular, data-driven approaches compute an approximation of the functional

$$\mathcal{S}(\arg\min_{v \in V} E(v, I_0, I_1)) \tag{5}$$

where $\mathcal{S}$ represents the operations needed to compute $(\phi_1^v)^{-1}$ from v, and the energy E is either given by Eqs. 1 or 3. The functional approximation is obtained via a neural network representation with parameters learned from a representative sample of image pairs. Unsupervised approaches assume that the LDDMM parameterization in combination with the minimization of the energy E considered as a loss function are enough for the inference of suitable diffeomorphic transformations after training. Therefore, there is no need for ground truth deformations. GAN-based approaches depart from unsupervised approaches by the definition of two different networks: the generative network (G) and the discrimination network (D), and are trained in an adversarial fashion as follows. The discrimination network D learns to distinguish between a warped source image $I_0 \circ (\phi_1^v)^{-1}$ generated by G and a plausible warped source image. It is trained using the loss function

$$L_D = \begin{cases} -\log(p) & c \in P^+ \\ -\log(1-p) & c \in P^- \end{cases} \tag{6}$$

where c indicates the input case, P^+ and P^- indicate positive or negative cases for the GAN, and p is the probability computed by D for the input case. In the first place, D is trained on a positive case $c \in P^+$ representing a target image I_1 and a warped source image I_0^w plausibly registered to I_1 with a diffeomorphic transformation. The warped source image is modeled from I_0 and I_1 with a strictly convex linear combination: $I_0^w = \beta I_0 + (1-\beta)I_1$. It should be noticed that, although the warped source image would ideally be I_1, the selection of $I_0^w = I_1$ (e.g. $\beta = 0$) empirically leads to the discriminator rapidly outperforming the generator. This approach to discriminators has been successfully used in adversarial learning methods for deformable registration [25]. Next, D is trained on a negative case $c \in P^-$ representing a target image I_1 and a warped source image I_0^w obtained from the generator network G. The generative network in this context is the diffeomorphic registration network. G is aimed at the approximation of the functional given in Eq. 5 similarly to unsupervised approaches for the inference of $(\phi_1^v)^{-1}$. It is trained using the combined loss function

$$L_G = L_{\text{adv}} + \lambda E(v, I_0, I_1). \tag{7}$$

where L_{adv} is the adversarial loss function, defined from $L_{\text{adv}} = -\log(p)$ where p is computed from D; E is the LDDMM energy given by Eqs. 1 or 3; and λ is the weight for balancing the adversarial and the generative losses. For each sample pair (I_0^w, I_1), G is fed with the pair of images and updates the network parameters from the back-propagation of the information of the loss function values coming from the LDDMM energy and the discriminator probability of being a pair generated by G.

3.1 Proposed GAN Architecture

Generator Network. In this work, the diffeomorphic registration network G is intended to learn LDDMM diffeomorphic registration parameterized on the space of steady velocity fields or the space of initial velocity fields subject to the EPDiff equation (Eq. 4). The diffeomorphic transformation $(\phi_1^v)^{-1}$ is obtained from these velocity fields either from scaling and squaring [7,8] or the solution of the deformation state equation [5]. Euler integration is used as PDE solver for all the involved differential equations. A number of different generator network architectures have been proposed in the recent literature, with predominance of simple fully convolutional (FC) [23] or U-Net like architectures [24,25]. In this work, we propose to use the architecture by Duan et al. [24] adapted to fit our purposes. The network follows the general U-net design of utilizing an encoder-decoder structure with skip connections. However, during the encoding phase, the source and target images are fed to two encoding streams with different resolution levels. The combination of the two encoding streams allows a larger receptive field suitable to learn large deformations. The upsampling is performed with a deconvolutional operation based on transposed convolutional layers [31]. We have empirically noticed that the learnable parameters of these layers help reduce typical checkerboard GAN artifacts in the decoding [32].

Discriminator Network. The discriminator network D follows a traditional CNN architecture. The two input images are concatenated and passed through five convolutional blocks. Each block includes a convolutional layer, a RELU activation function, and a size-two max-pooling layer. After the convolutions, the 4D volume is flattened and passed through three fully connected layers. The output of the last layer is the probability of the input images to come from a registered pair not generated by G.

Generative-Discriminative Integration Layer. The generator and the discriminator networks G and D are connected through an integration layer. This integration layer allows calculating the diffeomorphism $(\phi_1^v)^{-1}$ that warps the source image I_0. The selected integration layer depends on the velocity parameterization: stationary (SVF-GAN) or EPDiff-constrained time-dependent (EPDiff-GAN). The computed diffeomorphisms are applied to the source image via a second 3D spatial transformation layer [33] with no learnable parameters.

Parameter Selection and Implementation Details We selected the parameters $\lambda = 1000$, $\sigma^2 = 1.0$, $\alpha = 0.0025$, and $s = 4$ and a unit-domain discretization of the image domain Ω [5]. Scaling and squaring and Euler integration were performed in 8 and 10 time samples respectively. The parameter β for the convex linear modeling of warped images was selected equal to 0.2. Both the generator network and the discriminator network were trained with Adam's optimizer with default parameters and learning rates of $5e^{-5}$ for G and $1e^{-6}$ for D, respectively. The experiments were run on a machine equipped with one NVidia Titan RTX with 24 GBS of video memory and an Intel Core i7 with 64 GBS of DDR3 RAM, and developed in Python with Keras and a TensorFlow backend.

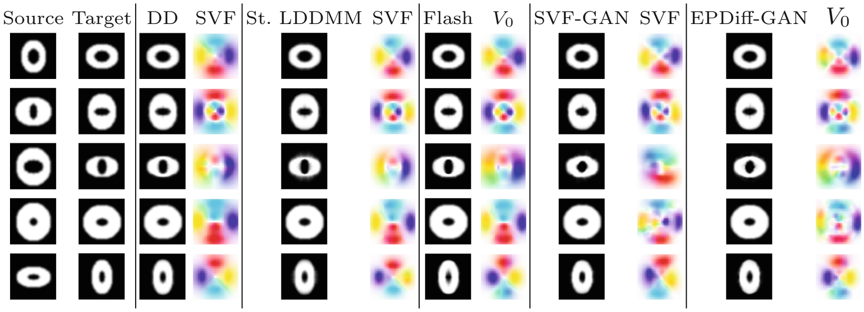


Fig. 1. Example of simulated 2D registration results. Up: source and target images of five selected experiments. Down, left to right: deformed images and velocity fields computed from diffeomorphic Demons (DD), stationary LDDMM (St. LDDMM), Flash, and our proposed SVF-GAN and EPDiff-GAN. SVF stands for a stationary velocity field and V_0 for the initial velocity field of a geodesic shooting approach, respectively.

4 Experiments and Results

2D Simulated Dataset. We simulated a total of 2560 torus images by varying the parameters of two ellipse equations, similarly to [19]. The parameters were drawn from two Gaussian distributions: $\mathcal{N}(4, 2)$ for the inner ellipse and $\mathcal{N}(12, 4)$ for the outer ellipse. The simulated images were of size 64 × 64. The networks were trained during 1000 epochs with a batch size of 64 samples.

3D Brain MRI Datasets. We used a total of 2113 T1-weighted brain MRI images from the Alzheimer's Disease Neuroimaging Initiative (ADNI). The images were acquired at the baseline visit and belong to all the available ADNI projects (1, 2, Go, and 3). The images were preprocessed with N3 bias field correction, affinely registered to the MNI152 atlas, skull-stripped, and affinely registered to the skull-stripped MNI152 atlas. The evaluation of our generated GAN models in the task of diffeomorphic registration was performed in NIREP dataset [34], where one image was chosen as reference and pair-wise registration was performed with the remaining 15. All images were scaled to size 176 × 224 × 176, and in this case trained for 50 epochs with a batch size of 1 sample. Inference of either a stationary or a time dependent velocity field takes 1.3 s.

Results in the 2D Simulated Dataset. Figure 1 show the deformed images and the velocity fields obtained in the 2D simulated dataset by diffeomorphic Demons [7], a stationary version of LDDMM (St. LDDMM) [8], the spatial version of Flash [10], and our proposed SVF and EPDiff GANs. Apart from diffeomorphic Demons that uses Gaussian smoothing for regularization, all the considered methods use the same parameters for operator L. Therefore, St. LDDMM and SVF-GAN can be seen as a model-based and a data-based approach for the minimization of the same variational problem. The same happens with Flash and EPDiff-GAN. From the figure, it can be appreciated that our proposed GANs are able to obtain accurate warps of the source to the target images, similarly to model-based approaches. For SVF-GAN, the inferred velocity fields are visually similar to model-based approaches in three of five experiments. For EPDiff-GAN, the inferred initial velocity fields are visually similar to model-based approaches in four of five experiments.

4.1 Results in the 3D NIREP Dataset

Quantitative Assessment. Figure 2 shows the Dice similarity coefficients obtained with diffeomorphic Demons [7], St. LDDMM [8], Voxelmorph II [16], the spatial version of Flash [10], Quicksilver [14] and our proposed SVF and EPDiff GANs. SVF-GAN shows an accuracy similar to St. LDDMM and competitive with diffeomorphic Demons. Our proposed method tends to overpass Voxelmorph II in the great majority of the structures. On the other hand, EPDiff-GAN shows an accuracy similar to Flash and Quicksilver in the great majority of regions, with the exception of the temporal pole (TP) and the orbital frontal gyrus (OFG), two small localized and difficult to register regions. Furthermore, the two-stream architecture greatly improves the accuracy obtained by a simple U-Net. SVF-GAN outperforms the ablation study model in which no discriminator was used, though EPDiff-GAN only shows clear performance improvements in some structures. It drives our attention that Flash underperformed in the superior frontal gyrus (SFG). All tested methods generate smooth deformations with almost no foldings, as can be seen in table 1 from the supplementary material.

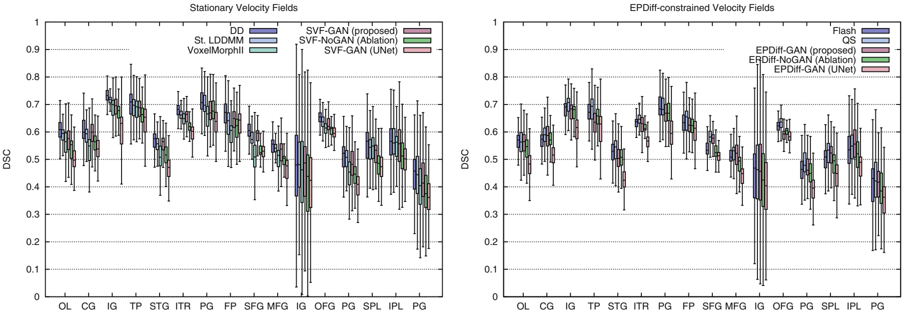


Fig. 2. Evaluation in NIREP. Dice scores obtained by propagating the diffeomorphisms to the segmentation labels on the 16 NIREP brain structures. Left, methods parameterized with stationary velocity fields: diffeomorphic Demons (DD), stationary LDDMM (St. LDDMM), Voxelmorph II, our proposed SVF-GAN with the two-stream architecture, SVF-GAN without discriminator and SVF-GAN with a U-net. Right, geodesic shooting methods: Flash, Quicksilver (QS), our proposed EPDiff-GAN, EPDiff-GAN without discriminator, and EPDiff-GAN with a U-net.

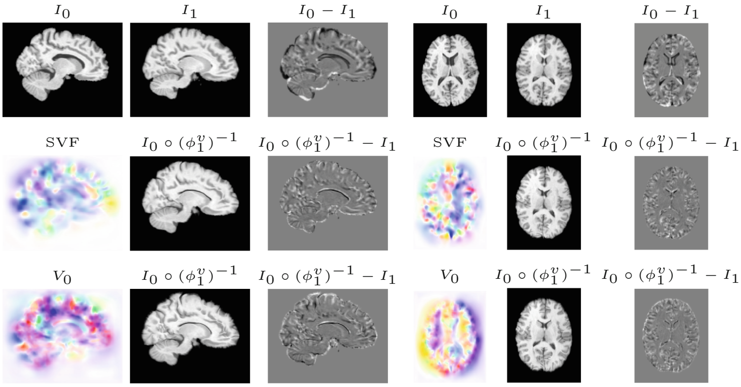

I_0
I_1
$I_0 - I_1$
I_0
I_1
$I_0 - I_1$
SVF
$I_0 \circ (\phi_1^v)^{-1}$
$I_0 \circ (\phi_1^v)^{-1} - I_1$
SVF
$I_0 \circ (\phi_1^v)^{-1}$
$I_0 \circ (\phi_1^v)^{-1} - I_1$
V_0
$I_0 \circ (\phi_1^v)^{-1}$
$I_0 \circ (\phi_1^v)^{-1} - I_1$
V_0
$I_0 \circ (\phi_1^v)^{-1}$
$I_0 \circ (\phi_1^v)^{-1} - I_1$

Fig. 3. Example of 3D registration results. First row, sagittal and axial views of the source and the target images and the differences before registration. Second row, inferred stationary velocity field, warped image, and differences after registration for SVF-GAN. Third row, inferred initial velocity field, warped image, and differences after registration for EPDiff-GAN.

Qualitative Assessment. For a qualitative assessment of the quality of the registration results, Fig. 3 shows the sagittal and axial views of one selected NIREP registration result. In the figure, it can be appreciated a high matching between the target and the warped ventricles, and more difficult to register regions like the cingulate gyrus (observable in the sagittal view) or the insular cortex (observable in the axial view).

5 Conclusions

We have proposed an adversarial learning LDDMM method for the registration of 3D mono-modal images. We have successfully implemented two models: one for the stationary parameterization and the other for the EPDiff-constrained non-stationary parameterization (geodesic shooting). The performed ablation study shows how GANs improve the results of the proposed registration networks. Furthermore, our experiments have shown that the inferred velocity fields are comparable to the solutions of model-based approaches. In addition, the evaluation study has shown the competitiveness of our approach with state of the art model- and data- based methods. It should be remarked that our methods perform similarly to Quicksilver, a supervised method that uses patches for training, and therefore, it learns in a rich-data environment. In contrast, our method is unsupervised and uses the whole image for training in a data-hungry environment. Indeed, our proposed methods outperform Voxelmorph II, an unsupervised method for diffeomorphic registration usually selected as benchmark in the state of the art. Finally, our proposal may constitute a good candidate for the massive computation of diffeomorphisms in Computational Anatomy studies, since once training has been completed, our method shows a computational time of over a second for the inference of velocity fields.

Acknowledgement. This work was partially supported by the national research grant TIN2016-80347-R (DIAMOND project),PID2019-104358RB-I00 (DL-Ageing project), and Government of Aragon Group Reference $T64\_20R$ (COSMOS research group). Ubaldo Ramon-Julvez's work was partially supported by an Aragon Government grant. Project PID2019-104358RB-I00 granted by MCIN/AEI/10.13039/501100011033. We would like to thank Gary Christensen for providing the access to the NIREP database [34]. Data used in the preparation of this article were partially obtained from the Alzheimer' s Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in the analysis or writing of this report. A complete listing of ADNI investigators can be found at: https://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

References

1. Sotiras, A., Davatzikos, C., Paragios, N.: Deformable medical image registration: a survey. IEEE Trans. Med. Imaging **32**(7), 1153–1190 (2013)
2. Modersitzki, J.: FAIR: Flexible Algorithms for Image Registration. SIAM, New Delhi (2009)
3. Hua, X.: ADNI: tensor-based morphometry as a neuroimaging biomarker for Alzheimer's disease: an MRI study of 676 AD, MCI, and normal subjects. Neuroimage **43**(3), 458–469 (2008)
4. Liu, Y., Li, Z., Ge, Q., Lin, N., Xiong, M.: Deep feature selection and causal analysis of Alzheimer's disease. Front. Neurosci. **13**, 1198 (2019)
5. Beg, M.F., Miller, M.I., Trouve, A., Younes, L.: Computing large deformation metric mappings via geodesic flows of diffeomorphisms. Int. J. Comput. Vision **61**(2), 139–157 (2005)
6. Ashburner, J.: A fast diffeomorphic image registration algorithm. Neuroimage **38**(1), 95–113 (2007)
7. Vercauteren, T., Pennec, X., Perchant, A., Ayache, N.: Diffeomorphic demons: efficient non-parametric image registration. Neuroimage **45**(1), S61–S72 (2009)
8. Hernandez, M.: Gauss-Newton inspired preconditioned optimization in large deformation diffeomorphic metric mapping. Phys. Med. Biol. **59**(20), 6805 (2014)
9. Vialard, F.X., Risser, L., Rueckert, D., Cotter, C.J.: Diffeomorphic 3D image registration via geodesic shooting using an efficient adjoint calculation. Int. J. Comput. Vision **97**(2), 229–241 (2011)
10. Zhang, M., Fletcher, T.: Fast diffeomorphic image registration via fourier-approximated lie algebras. Int. J. Comput. Vision **127**, 61–73 (2018)
11. Dosovitskiy, A., Fischere, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V.: Flownet: learning optical flow with convolutional networks. In: Proceedings of the 16th IEEE International Conference on Computer Vision (ICCV 2015), pp. 2758–2766 (2015)
12. Boveiri, H., Khayami, R., Javidan, R., Mehdizadeh, A.: Medical image registration using deep neural networks: a comprehensive review. Comput. Electr. Eng. **87**, 106767 (2020)

13. Rohé, M.-M., Datar, M., Heimann, T., Sermesant, M., Pennec, X.: SVF-Net: learning deformable image registration using shape matching. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10433, pp. 266–274. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66182-7_31
14. Yang, X., Kwitt, R., Styner, M., Niethammer, M.: Quicksilver: fast predictive image registration - a deep learning approach. Neuroimage **158**, 378–396 (2017)
15. Dalca, A.V., Balakrishnan, G., Guttag, J., Sabuncu, M.R.: Unsupervised learning for fast probabilistic diffeomorphic registration. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11070, pp. 729–738. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00928-1_82
16. 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)
17. Krebs, J., Delingetter, H., Mailhe, B., Ayache, N., Mansi, T.: Learning a probabilistic model for diffeomorphic registration. IEEE Trans. Med. Imaging **38**, 2165–2176 (2019)
18. Fan, J., Cao, X., Yap, P., Shen, D.: BIRNet: brain image registration using dual-supervised fully convolutional networks. Med. Image Anal. **54**, 193–206 (2019)
19. Wang, J., Zhang, M.: DeepFLASH: an efficient network for learning-based medical image registration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2020) (2020)
20. Mok, T.C.W., Chung, A.C.S.: Fast symmetric diffeomorphic image registration with convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2020) (2020)
21. Hoffmann, M., Billot, B., Greve, D.N., Iglesias, J.E., Fischl, B., Dalca, A.V.: Synthmorph: learning contrast-invariant registration without acquired images. IEEE Trans. Med. Imaging **41**(3), 543–558 (2021)
22. Dalca, A.V., Balakrishnan, G., Guttag, J., Sabuncu, M.R.: Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. Med. Image Anal. **57**, 226–236 (2019)
23. Mahapatra, D., Antony, B., Sedai, S., Garvani, R.: Deformable medical image registration using generative adversarial networks. In: IEEE International Symposium on Biomedical Imaging (ISBI 2018) (2018)
24. Duan, L., et al.: Adversarial learning for deformable registration of brain MR image using a multi-scale fully convolutional network. Biomed. Signal Process. Control **53**, 101562 (2018)
25. Fan, J., Cao, X., Wang, Q., Yap, P., Shen, D.: Adversarial learning for mono- or multi-modal registration. Med. Image Anal. **58**, 1015–1045 (2019)
26. Dey, N., Ren, M., Dalca, A.V., Gerig, G.: Generative adversarial registration for improved conditional deformable templates. In: Proceedings of the 18th IEEE International Conference on Computer Vision (ICCV 2021) (2021)
27. Dalca, A.V., Rakic, M., Guttag, J.V., Sabuncu, M.R.: Learning conditional deformable templates with convolutional networks. In: NeurIPS (2019)
28. Bigolin Lanfredi, R., Schroeder, J.D., Vachet, C., Tasdizen, T.: Interpretation of disease evidence for medical images using adversarial deformation fields. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12262, pp. 738–748. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59713-9_71
29. Younes, L.: Jacobi fields in groups of diffeomorphisms and applications. Q. Appl. Math. **65**, 113–134 (2007)

30. Arsigny, V., Commonwick, O., Pennec, X., Ayache, N.: Statistics on diffeomorphisms in a Log-Euclidean framework. In: Proceedings of the 9th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2006), Lecture Notes in Computer Science, vol. 4190, pp. 924–931 (2006)
31. Zeiler, M.D., Taylor, G.W., Fergus, R.: Adaptive deconvolutional networks for mid and high level feature learning. In: ICCV, vol. 2011, pp. 2018–2025 (2011)
32. Odena, A., Dumoulin, V., Olah, C.: Deconvolution and checkerboard artifacts. Distill **1**(10), e3 (2016)
33. Jaderberg, M., Simonyan, K., Zissermann, A., Kavukcuoglu, K.: Spatial transformer networks. In: Proceedings of Conference on Neural Information Processing Systems (NeurIPS 2015) (2015)
34. Christensen, G.E., et al.: Introduction to the non-rigid image registration evaluation project (NIREP). In: Proceedings of 3rd International Workshop on Biomedical Image Registration (WBIR 2006), vol. 4057, pp. 128–135 (2006)

Towards a 4D Spatio-Temporal Atlas of the Embryonic and Fetal Brain Using a Deep Learning Approach for Groupwise Image Registration

Wietske A. P. Bastiaansen[1,2](✉), Melek Rousian[2],
Régine P. M. Steegers-Theunissen[2], Wiro J. Niessen[1], Anton H. J. Koning[3],
and Stefan Klein[1]

[1] Department of Radiology and Nuclear Medicine, Biomedical Imaging Group Rotterdam, Erasmus MC, Rotterdam, Netherlands
w.bastiaansen@erasmusmc.nl

[2] Department of Obstetrics and Gynecology, Erasmus MC, Rotterdam, Netherlands

[3] Department of Pathology, Erasmus MC, Rotterdam, Netherlands

Abstract. Brain development during the first trimester is of crucial importance for current and future health of the fetus, and therefore the availability of a spatio-temporal atlas would lead to more in-depth insight into the growth and development during this period. Here, we propose a deep learning approach for creation of a 4D spatio-temporal atlas of the embryonic and fetal brain using groupwise image registration. We build on top of the extension of Voxelmorph for the creation of learned conditional atlases, which consists of an atlas generation and registration network. As a preliminary experiment we trained only the registration network and iteratively updated the atlas. Three-dimensional ultrasound data acquired between the 8th and 12th week of pregnancy were used. We found that in the atlas several relevant brain structures were visible. In future work the atlas generation network will be incorporated and we will further explore, using the atlas, correlations between maternal periconceptional health and brain growth and development.

Keywords: Embryonic and fetal brain atlas · Groupwise image registration · First trimester ultrasound · Deep learning

1 Introduction

Normal growth and development of the human embryonic and fetal brain during the first trimester is of crucial importance for current and future health of the fetus [14,17]. Currently, this is monitored by manual measurements, such as the circumference and volume of the brain [13,15]. However, these measurements lack

A. Hering et al. (Eds.): WBIR 2022, LNCS 13386, pp. 29–34, 2022.
https://doi.org/10.1007/978-3-031-11203-4_4

overview: it is unclear how the different measurements relate. The availability of an atlas i.e., a set of brain templates for a range of gestational ages, could overcome these challenges by offering a unified and automatic framework to compare development across subjects.

In literature several atlases are available [5–8,11,12,16,18,19]. However, these are based on magnetic resonance imaging and/or acquired during the second and third trimester of pregnancy. Here, we present to the best of our knowledge the first framework for the development of a brain atlas describing growth of the human embryo and fetus between 56 and 90 days gestational age (GA) based on ultrasound imaging.

2 Method

The atlas is generated from three-dimensional (3D) ultrasound images $I_{i,t}$, for subject i imaged at time t, where t is the GA in days. The atlas A_t is obtained by groupwise registration of $I_{i,t}$ for every pregnancy $i = 1, ..., k$ on every time point t, followed by taking the mean over the deformed images: $A_t = \frac{1}{k} \sum_i I_{i,t} \circ \phi_i$. Hereby the constraint $\sum_{i,t} \phi_{i,t} \approx 0$ is applied, as proposed by Balci et al. and Bhatia et al. [1,3]. To ensure invertibility of the deformations we used diffeomorphic non-rigid deformations with the deformation field $\phi_{i,t}$, obtained by integrating the velocity field $\nu_{i,t}$.

The framework is based on the extension of Voxelmorph for learning conditional atlases by Dalca et al. [4]. An overview of the framework can be found in Fig. 1. Here, we only train the registration framework and we initialize the atlas for every time t as the voxelwise median over all images $I_i \forall i$. The median was chosen over the mean, since this resulted in a sharper initial atlas. Next, the atlas is updated for iteration n as the mean of $I_{i,t} \circ \phi^n_{i,t}$ for every time t. Subsequently, the network is trained until $A^n_t \approx A^{n-1}_t$.

The loss function is defined as follows:

$$\begin{aligned}\mathcal{L}\left(A_t, I_{i,t}, \phi_{i,t}, \phi^{-1}_{i,t}\right) &= \lambda_{\text{sim}} \mathcal{L}_{\text{similarity}}\left(A_t \circ \phi^{-1}_{i,t}, I_{i,t}\right) + \lambda_{\text{group}} \mathcal{L}_{\text{groupwise}}\left(\phi_{i,t}\right) \\ &+ \lambda_{\text{mag}} \mathcal{L}_{\text{magnitude}}\left(\phi^{-1}_{i,t}\right) + \lambda_{\text{dif}} \mathcal{L}_{\text{diffusion}}\left(\phi^{-1}_{i,t}\right)\end{aligned} \tag{1}$$

The first term computes the similarly between the atlas and image, we used the local squared normalized cross-correlation, which was used before on this dataset [2]. The second term approximates the constraint for groupwise registration by minimizing the running average over the last c deformation fields obtained during training. To balance the influence of this constraint with respect to time, we sorted the data based on day GA within every epoch and took as window c the average number of images per day GA in the dataset. Finally, the deformations are regularized by: $\mathcal{L}_{\text{mag}} = \|\phi^{-1}_{i,t}\|^2_2$ and $\mathcal{L}_{\text{dif}} = \|\nabla\phi^{-1}_{i,t}\|^2_2$.

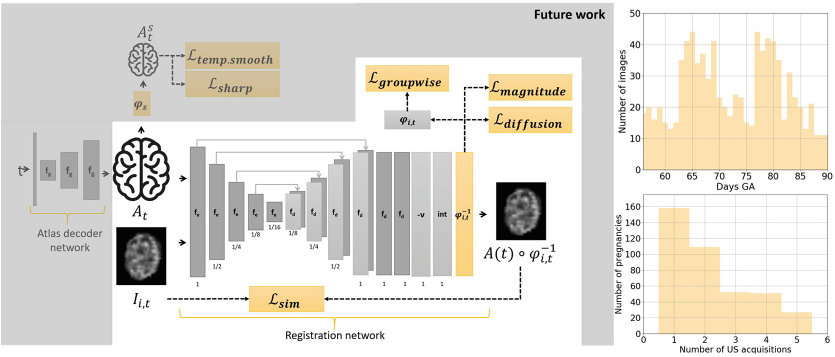


Fig. 1. Overview of the proposed framework and characteristics of the used dataset.

3 Data and Experiments

The Rotterdam Periconceptional Cohort (Predict study) is a large hospital-based cohort study conducted at the Erasmus MC, University Medical Center Rotterdam, the Netherlands. This prospective cohort focuses on the relationships between periconceptional maternal and paternal health and embryonic and fetal growth and development [14,17]. 3D ultrasound scans are acquired at multiple points in time during the first trimester. Here, to model normal development, we included only singleton pregnancies with no adverse outcome and spontaneous conception with a regular menstrual cycle.

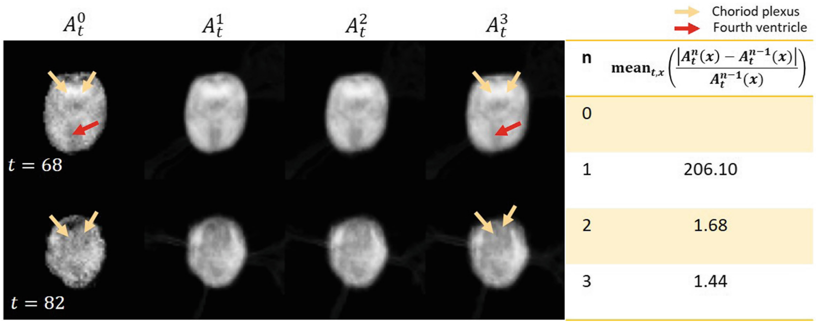
We included 871 ultrasound images of 398 pregnancies acquired between 56 and 90 days GA. For each day GA, we have at least 10 ultrasound images, as shown in top-right graph in Fig. 1. The data was split such that for every day GA 80% of the data is in the training set and 20% in the test set. We first spatially aligned and segmented the brain using our previously developed algorithm for multi-atlas segmentation and registration of the embryo [2]. Next, we resized all images to a standard voxelsize per day GA, to ensure that the brain always filled a similar field of view despite the fast growth of the brain. This standard voxelsize per day GA was determined by linear interpolation of the median voxelsize per week GA. We trained the network using the default hyperparameters proposed by Dalca et al. [4] for $\lambda_{\text{group}} \in \{0, 1, 10, 100\}$. We reported the mean percentage of voxels having a non-positive Jacobian determinant $\%|J| \leq 0$, the groupwise loss $\mathcal{L}_{\text{group}}$ and the similarity loss $\mathcal{L}_{\text{sim}}$. Finally, for the best set of hyperparameters the atlas was updated iteratively, and we visually analyzed the result.

4 Results

From the results given in Table 1 for iteration $n = 1$ we concluded that all tested hyperparameters resulted in smooth deformation fields, since the percentage of voxels with a non-positive Jacobian determinant $\%|J| \leq 0$ over the whole dataset was less then one percent. Furthermore, we observe that for $\lambda_{\text{group}} = 1$ $\mathcal{L}_{\text{group}}$ is similar to not enforcing the groupwise constraint. For $\lambda_{\text{group}} = 100$, we observed that $\mathcal{L}_{\text{sim}}$ deteriorated, indicating that the deformation fields are excessively restricted by the groupwise constraint. Hence, $\lambda_{\text{group}} = 10$ was used to iteratively update the atlas. Finally, note that the difference between results for training and testing are minimal: indicating a limited degree of overfitting. In Fig. 2 a visualization of the results can be found for $t = 68$ and $t = 82$. In the showed axial slices the choroid plexus and the fourth ventricle can be observed.

Table 1. Results for different hyperparameters, with the standard deviation given between brackets.

Hyperparameters				Training			Test		
λ_{sim}	λ_{group}	λ_{mag}	λ_{dif}	$\%\lvert J\rvert \leq 0$	$\mathcal{L}_{\text{group}}$	$\mathcal{L}_{\text{sim}}$	$\%\lvert J\rvert \leq 0$	$\mathcal{L}_{\text{group}}$	$\mathcal{L}_{\text{sim}}$
1	0	0.01	0.01	0.26 (0.47)	1.45e−3	0.126	0.36 (0.55)	1.90e−3	0.130
1	1	0.01	0.01	0.25 (0.38)	1.27e−3	0.125	0.32 (0.42)	1.62e−3	0.129
1	10	0.01	0.01	0.17 (0.27)	7.80e−4	0.118	0.22 (0.28)	9.40e−4	0.126
1	100	0.01	0.01	0.04 (0.06)	1.61e−4	0.091	0.05 (0.07)	1.85e−4	0.101



n	$\text{mean}_{t,x}\left(\frac{\lvert A_t^n(x) - A_t^{n-1}(x)\rvert}{A_t^{n-1}(x)}\right)$
0	
1	206.10
2	1.68
3	1.44

Fig. 2. Axial slice of the atlas for different GA and iterations 0, 1, 2 and 3.

5 Discussion and Conclusion

We propose a deep learning approach for creation of a 4D spatio-temporal atlas of the embryonic and fetal brain using groupwise image registration. Here, we trained the registration network iteratively and visually inspected the resulting atlas. We found that the registration network results in smooth deformation field, and that several relevant brain structures were visible in the atlas.

In this work, the window c of the groupwise loss term was set to the mean number of samples per day GA, in future work this hyperparameter will be varied to study its influence. As shown in Fig. 1, in future work also the atlas generator network will be incorporated, where constraints for temporal smoothness and sharp edges in the atlas can directly be incorporated in the loss. Finally, we will evaluate if the relevant brain measurements of the atlas are close to clinically known values and we will analyze if the morphology of the brain, modelled by the deformations $\phi_{i,t}$, shows the known correlation with maternal periconceptional health factors found in previous research [9,10].

References

1. Balci, S.K., Golland, P., Shenton, M.E., Wells, W.M.: Free-form B-spline deformation model for groupwise registration. Med. Image Comput. Comput. Assist. Interv. **10**, 23–30 (2007)
2. Bastiaansen, W.A., Rousian, M., Steegers-Theunissen, R.P., Niessen, W.J., Koning, A.H., Klein, S.: Multi-atlas segmentation and spatial alignment of the human embryo in first trimester 3D ultrasound. arXiv:2202.06599 (2022)
3. Bhatia, K., Hajnal, J., Puri, B., Edwards, A., Rueckert, D.: Consistent groupwise non-rigid registration for atlas construction. In: 2004 2nd IEEE International Symposium on Biomedical Imaging: Nano to Macro, vol. 1, pp. 908–911 (2004)
4. Dalca, A., Rakic, M., Guttag, J., Sabuncu, M.: Learning conditional deformable templates with convolutional networks. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
5. Dittrich, E., et al.: A spatio-temporal latent atlas for semi-supervised learning of fetal brain segmentations and morphological age estimation. Med. Image Anal. **18**(1), 9–21 (2014)
6. Gholipour, A.: A normative spatiotemporal MRI atlas of the fetal brain for automatic segmentation and analysis of early brain growth. Sci. Rep. **7**(1), 1–13 (2017)
7. Habas, P.A., et al.: A spatiotemporal atlas of MR intensity, tissue probability and shape of the fetal brain with application to segmentation. Neuroimage **53**(2), 460–470 (2010)
8. Khan, S., et al.: Fetal brain growth portrayed by a spatiotemporal diffusion tensor MRI atlas computed from in utero images. Neuroimage **185**, 593–608 (2019)
9. Koning, I., et al.: Growth trajectories of the human embryonic head and periconceptional maternal conditions. Hum. Reprod. **31**(5), 968–976 (2016)
10. Koning, I., Dudink, J., Groenenberg, I., Willemsen, S., Reiss, I., Steegers-Theunissen, R.: Prenatal cerebellar growth trajectories and the impact of periconceptional maternal and fetal factors. Hum. Reprod. **32**(6), 1230–1237 (2017)
11. Kuklisova-Murgasova, M., et al.: A dynamic 4D probabilistic atlas of the developing brain. Neuroimage **54**(4), 2750–2763 (2011)
12. Namburete, A.I.L., van Kampen, R., Papageorghiou, A.T., Papież, B.W.: Multichannel groupwise registration to construct an ultrasound-specific fetal brain atlas. In: Melbourne, A., et al. (eds.) PIPPI/DATRA -2018. LNCS, vol. 11076, pp. 76–86. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00807-9_8
13. Paladini, D., Malinger, G., Birnbaum, R., Monteagudo, A., Pilu, G., Salomon, L.: ISUOG practice guidelines (updated): sonographic examination of the fetal central nervous system. Part 1: performance of screening examination and indications for targeted neurosonography. Ultrasound Obstet. Gynecol. **56**, 476–484 (2020)

14. Rousian, M., et al.: Cohort profile update: the Rotterdam Periconceptional Cohort and embryonic and fetal measurements using 3D ultrasound and virtual reality techniques. Int. J. Epidemiol. **50**, 1–14 (2021)
15. Salomon, L.J., et al.: Practice guidelines for performance of the routine mid-trimester fetal ultrasound scan. Ultrasound Obstet. Gynecol. **37**(1), 116–126 (2011)
16. Serag, A., et al.: Construction of a consistent high-definition Spatio-temporal atlas of the developing brain using adaptive kernel regression. Neuroimage **59**(3), 2255–2265 (2012)
17. Steegers-Theunissen, R., et al.: Cohort profile: the Rotterdam Periconceptional cohort (predict study). Int. J. Epidemiol. **45**, 374–381 (2016)
18. Uus, A., et al.: Multi-channel 4D parametrized atlas of macro-and microstructural neonatal brain development. Frontiers Neurosci., 721 (2021)
19. Wu, J., et al.: Age-specific structural fetal brain atlases construction and cortical development quantification for Chinese population. Neuroimage **241**, 118412 (2021)

Uncertainty

DeepSTAPLE: Learning to Predict Multimodal Registration Quality for Unsupervised Domain Adaptation

Christian Weihsbach(✉), Alexander Bigalke, Christian N. Kruse, Hellena Hempe, and Mattias P. Heinrich

Institute of Medical Informatics, Universität zu Lübeck,
Ratzeburger Allee 160, 23538 Lübeck, Germany
christian.weihsbach@uni-luebeck.de
https://www.imi.uni-luebeck.de/en/institute.html

Abstract. While deep neural networks often achieve outstanding results on semantic segmentation tasks within a dataset domain, performance can drop significantly when predicting domain-shifted input data. Multi-atlas segmentation utilizes multiple available sample annotations which are deformed and propagated to the target domain via multimodal image registration and fused to a consensus label afterwards but subsequent network training with the registered data may not yield optimal results due to registration errors. In this work, we propose to extend a curriculum learning approach with additional regularization and fixed weighting to train a semantic segmentation model along with data parameters representing the atlas confidence. Using these adjustments we can show that registration quality information can be extracted out of a semantic segmentation model and further be used to create label consensi when using a straightforward weighting scheme. Comparing our results to the STAPLE method, we find that our consensi are not only a better approximation of the oracle-label regarding Dice score but also improve subsequent network training results.

Keywords: Domain adaptation · Multi-atlas registration · Label noise · Consensus · Curriculum learning

1 Introduction

Deep neural networks dominate the state-of-the-art medical image segmentation [10,14,20], but their high performance is depending on the availability of large-scale labelled datasets. Such labelled data is often not available in the target domain and direct transfer learning leads to performance drops due to domain shift [27]. To overcome these issues transferring existing annotations from a labeled source to the target domain is desirable. Mutli-atlas segmentation is a popular method, which accomplishes such a label transfer in two steps: First,

A. Hering et al. (Eds.): WBIR 2022, LNCS 13386, pp. 37–46, 2022.
https://doi.org/10.1007/978-3-031-11203-4_5

multiple sample annotations are transferred to target images via image registration [7,18,24] resulting in multiple "optimal" labels [1]. Secondly label fusion can be applied to build the label consensus. Although many methods for finding a consensus label have been developed [1,6,19,25,26], the resulting fused labels are still not perfect and exhibit label noise, which complicates the training of neural networks and degrades performance.

Related Work. In the past, various label fusion methods have been proposed, which use weighted voting on registered label candidates to output a common consensus label [1,6,19,26]. More elaborate fusion methods also use image intensities [25], however when predicting across domains source and target intensities can differ substantially complicating intensity-based fusion and would therefore require handling of the intensity gap i.e. with image-to-image translation techniques [29]. When using the resulting consensus labels from non-optimal registration and fusion for subsequent CNN training, noisy data is introduced to the network [12]. Network training can then be improved with techniques of curriculum learning to estimate label noise (i.e. difficulty) and guide the optimization process accordingly [3,22] but the techniques have not been used in the context of noise introduced through registered pixel-wise labels [2,3,11,22,28] or employ more specialized and complex pipelines [4,5,15]. Other deep learning-based techniques to address ambiguous labels are probabilistic networks [13].

Contributions. We propose to use data parameters [22] to weight noisy atlas samples as a simple but effective extension of semantic segmentation models. During training the data parameters (scalar values assigned to each instance of a registered label) can estimate the label trustworthiness globally across all multi-atlas candidates of all images. We extend the original formulation of data parameters by additional *risk regularization* and *fixed weighting* terms to adapt to the specific characteristics of the segmentation task and show that our adaptation improves network training performance for 2D and 3D tasks in the single-atlas scenario. Furthermore, we apply our method to the multi-atlas 3D image scenario where the network scores do not improve but yield equal performance in comparison to normal cross-entropy loss training when using out-of-line backpropagation. Nonetheless, we still can achieve an improvement by deriving an optimized consensus label from the extracted weights and applying a straightforward weighted-sum on the registered atlases.

2 Method

In this section, we will describe our data parameter adaption[1] and introduce our proposed extensions when using it in semantic segmentation tasks, namely a special regularization and a fixed weighting scheme. Furthermore, a multi-atlas specific extension will be described, which improves training stability.

[1] Our code is openly available on GitHub: https://github.com/multimodallearning/deep_staple.

Data Parameters. Saxena et al. [22] formulate their data parameter and curriculum learning approach as a modification altering the logits input of the loss function. By a learnable logits-weighting improvements could be shown in different scenarios when either noisy training samples and/or classes were weighted during training. Our implementation and experiments focus on per-sample parameters $\mathbf{DP_S}$ of a dataset $S = \{(\mathbf{x_s}, \mathbf{y_s})\}_{s=1}^{n}$ with images x_s and labels y_s containing n training samples. Since weighting schemes for multi-atlas label fusion like STAPLE [26] use a confidence weight of 0 indicating "no confidence" and 1 indicating "maximum confidence we slightly changed the initial formulation of data parameters:

$$\mathbf{DP}_{\sigma} = \mathit{sigmoid}\left(\mathbf{DP_S}\right) \tag{1}$$

According to Eq. 1 we limit the data parameters applied to our loss to $DP_{\sigma} \in (0, 1)$ where a value of 0 indicates "no confidence" and 1 indicates "maximum confidence" such as weighting schemes like STAPLE [26]. The data parameter loss ℓ_{DP} is calculated as

$$\ell_{DP}\left(f_{\theta}\left(\mathbf{x_B}\right), \mathbf{y_B}\right) = \sum_{b=1}^{|B|} \ell_{CE,spatial}\left(f_{\theta}\left(\mathbf{x_b}\right), \mathbf{y_b}\right) \cdot DP_{\sigma_b} \quad \text{with} \quad B \subseteq S \tag{2}$$

where B is a training batch, $\ell_{CE,spatial}$ is the cross-entropy loss reduced over spatial dimensions and f_{θ} the model. As in the original implementation, the parameters require a sparse implementation of the Adam optimizer to avoid diminishing momenta. Note, that the data parameter layer is omitted for inference—inference scores are only affected indirectly by data parameters through optimized model training.

Risk Regularisation. Even when a foreground class is present in the image and a registered target label only contains background voxels, the network can achieve a zero-loss value by overfitting. As a consequence, upweighting the overfitted samples will be of no harm in terms of loss reduction which leads to the upweighting of maximal noisy (empty) samples. We therefore add a so called *risk regularisation* encouraging the network to take *risk*

$$\ell = \ell_{DP} - \sum_{b=1}^{|B|} \frac{\#\left\{f_{\theta}\left(\mathbf{x_b}\right) = c\right\}}{\#\left\{f_{\theta}\left(\mathbf{x_b}\right) = c\right\} + \#\left\{f_{\theta}\left(\mathbf{x_b}\right) = \overline{c}\right\}} \cdot DP_{\sigma_b} \tag{3}$$

where $\#\left\{f_{\theta}\left(\mathbf{x_b}\right) = c\right\}$ and $\#\left\{f_{\theta}\left(\mathbf{x_b}\right) = \overline{c}\right\}$ indicate positive and negative predicted voxel count. According to this regularisation the network can reduce loss when predicting more target voxels under the restriction that the sample has a high data parameter value i.e. is classified as a clean sample. This formulation is balanced because predicting more positive voxels will increase the cross-entropy term if the prediction is inaccurate.

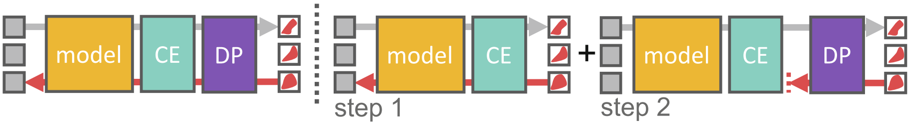


Fig. 1. Left: Inline backpropagation updating (red arrow) model and data parameters together. **Right:** Out-of-line backpropagation first steps on model (gray arrow) using normal cross-entropy loss and then steps on data parameters using the model's weights of the first step. (Color figure online)

Fixed Weighting Scheme. We found that the parameters have a strong correlation with the ground-truth voxels present in their values. Applying a fixed compensation weighting to the data parameters DP_{σ_b} can improve the correlation of the learned parameters and our target scores

$$DP_{\tilde{\sigma}_b} = \frac{DP_{\sigma_b}}{log\left(\#\left\{(\mathbf{y_b} = c\right\} + e\right) + e} \tag{4}$$

where $\#\left\{\mathbf{y_b} = c\right\}$ denotes the count of ground-truth voxels and e Euler's number.

Out-of-Line Backpropagation Process for Improved Stability. The interdependency of data parameters and model parameters can cause convergence issues when training *inline*, especially during earlier epochs when predictions are inaccurate. We found that a two-step forward-backward pass, first through the main model and in the second step through the main model and the data parameters can maintain stability while still estimating label noise (see Fig. 1). First only the main model parameters will be optimized. Secondly only the data parameters will be optimized *out-of-line*. When using the *out-of-line*, two-step approach data parameter optimization becomes a hypothesis of *"what would help the model optimizing right now?"* without intervening. Due to the optimizer momentum the parameter values still become reasonably separated.

Consensus Generation via Weighted Voting. To create a consensus $\mathbf{C_M}$ we use a simple weighted-sum over a set of multi-atlas labels M associated to a fixed image that turned out to be effective

$$\mathbf{C_M} = \left(\sum_{m=1}^{|M|} softmax(\mathbf{DP_M})_m \cdot \mathbf{y_m}\right) > 0.5 \quad \text{with} \quad M \subset S \tag{5}$$

where $\mathbf{DP_M}$ are the parameters associated to the set of multi-atlas labels $\mathbf{y_M}$.

3 Experiments

In this section, we will describe general dataset and model properties as well as our four experiments which increase in complexity up to the successful application of our method in 3D multi-atlas label noise estimation. We will refer to oracle-labels[2] as the real target labels which belong to an image and "registered/training/ground-truth"-labels as image labels that the network used to update its weights. Oracle-Dice refers to the overlapping area of oracle-labels and "registered/training/ground-truth"-labels.

Dataset. For our experiments, we chose a challenging multimodal segmentation task which was part of the CrossMoDa challenge [23]. The data contains contrast-enhanced T1-weighted brain tumour MRI scans and high-resolution T2-weighted images (initial resolution of $384/448 \times 348/448 \times 80$ *vox* @ $0.5\,\text{mm} \times 0.5\,\text{mm} \times 1.0{-}1.5\,\text{mm}$ and $512 \times 512 \times 120$ *vox* @ $0.4 \times 0.4 \times 1.0{-}1.5\,\text{mm}$). We used the original TCIA dataset [23] to provide omitted labels of the CrossModa challenge which served as oracle-labels. Prior to training isotropic resampling to $0.5\,\text{mm} \times 0.5\,\text{mm} \times 0.5\,\text{mm}$ was performed as well as cropping the data to $128 \times 128 \times 128$ *vox* around the tumour. We omitted the provided cochlea labels and train on binary masks of background/tumour. As the tumour is either contained on the right- or left side of the hemisphere, we flipped the right samples to provide pre-oriented training data and omit the data without tumour structures. For the 2D experiments we sliced the last data dimension.

Model and Training Settings. For 2D segmentation, we employ a LR-ASPP MobileNetV3-Large model [9]. For 3D experiments we use a custom 3D-MobileNet backbone similar as proposed in [21] with an adapted 3D-LR-ASPP head [8]. 2D training was performed with an AdamW [17] optimizer with a learning rate of $\lambda_{2D} = 0.0005$, $|B|_{2D} = 32$, cosine annealing [16] as scheduling method with restart after $t_0 = 500$ batch steps and multiplication factor of 2.0. For the data parameters, we used the SparseAdam-optimizer implementation together with the sparse Embedding structure of PyTorch with a learning rate of $\lambda_{DP} = 0.1$, no scheduling, $\beta_1 = 0.9$ and $\beta_2 = 0.999$. 3D training was conducted with learning rate of $\lambda_{3D} = 0.01$, $|B|_{3D} = 8$ due to memory restrictions and exponentially decayed scheduling with factor of $d = 0.99$. As opposed to Saxena et al. [22] during our experiments we did not find weight-clipping, weight decay or ℓ_2-regularisation on data parameters to be necessary. Parameters DP_s were initialized with a value of 0.0. For all experiments, we used spatial affine- and b-spline-augmentation and random-noise-augmentation on image intensities. Prior to augmenting we upscaled the input images and labels to 256×256 *px* in 2D- and $192 \times 192 \times 192$ *vox* in 3D-training. Data was split into 2/3 training and 1/3 validation images during all runs and used global class weights $1/{n_{bins}}^{0.35}$.

[2] "The word oracle [...] properly refers to the priest or priestess uttering the prediction.". "Oracle." Wikipedia, Wikimedia Foundation, 03 Feb 2022, en.wikipedia.org/wiki/Oracle.

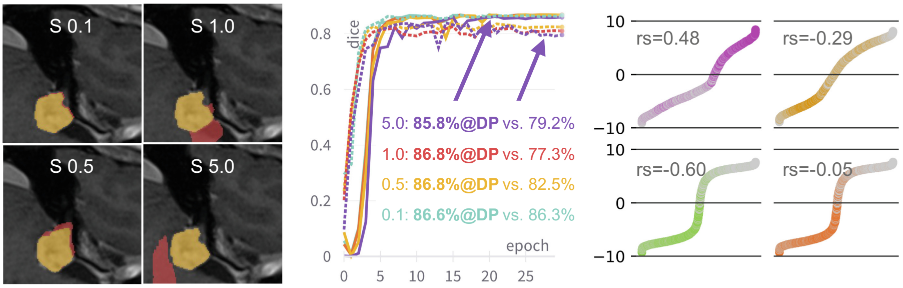


Fig. 2. Left: Sample disturbance 🟥 at strengths [0.1, 0.5, 1.0, 5.0]. **Middle:** Validation Dice when training with named disturbance strengths, either with data parameters enabled (—) or disabled (- -). **Right:** Parameter distribution for combinations of risk regularization (RR) and fixed weighting (FW): RR+FW 🟪 | RR 🟧 | FW 🟨 | NONE 🟩. Saturated data points indicate higher oracle-Dice. Value of ranked Spearman-correlation r_s between data parameters and oracle-Dice given. (Color figure online)

Experiment I: 2D Model Training, Artificially Disturbed Ground-Truth Labels. This experiment shows the general applicability of data parameters in the semantic segmentation setting when using one parameter per 2D slice. To simulate label-noise, we shifted 30% of the non-empty oracle-slices with different strengths (Fig. 2, left) to see how the network scores behave (Fig. 2, middle) and whether the data parameter distribution captures the artificially disturbed samples (Fig. 2, right). In case of runs with data parameters the optimization was enabled after 10 epochs.

Experiment II: 2D Model Training, Quality-Mixed Registered Single-Atlas Labels. Extending experiment I, in this setting we train on real registration noise with 2D slices on single-atlases. We use 30 T1-weighted images as fixed targets (non-labelled) and T2-weighted images and labels as moving pairs. For registration we use the deep learning-based algorithm Convex Adam [24]. We select two registration qualities to show quality influence during training: *Best*-quality registration means the single best registration with an average of around 80% oracle-Dice across all atlas registrations. *Combined*-quality means a clipped, gaussian-blurred sum of all 30 registered atlas registrations (some sort of consensus). We then input a mix of 50%/50% randomly selected best/combined labels into training. Afterwards we compare the 100% best, 50%/50% mixed and 100% combined selections focusing on the mixed setting where we train with and without data parameters. Validation scores were as follows (descending): best@no-data-parameters 81.1%, mix@data-parameters 74.1%, mix@no-data-parameters 69.6% and combined@no-data-parameters 61.9%.

Experiment III: 3D Model Training, Registered Multi-atlas Labels. Extending experiment II, in this setting we train on real registration noise but with 3D volumes and multiple atlases per image. We follow the CrossMoDa [23] challenge task and use T2-weighted images as fixed targets (non-labelled) and

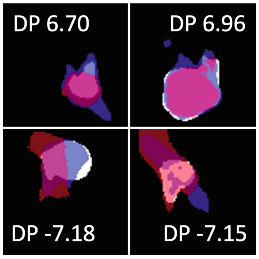


Fig. 3. Selected samples with low- and high parameters: Oracle-label □, network prediction 🟦 and deeds registered label 🟪 (Color figure online)

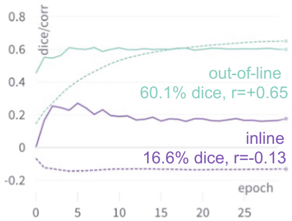


Fig. 4. Inline 🟪 and out-of-line 🟩 backpropagation. Validation Dice (—) and Spearman-corr. of params. and oracle-Dice (- -) (Color figure online)

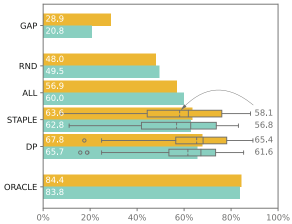


Fig. 5. FG: Box plots of STAPLE and DP consensus quality, mean value on the right. **BG:** Bar plot of nnUNet scores; deeds 🟩, Convex Adam 🟨 (Color figure online)

T1-weighted images and labels as moving pairs. We conducted registration with two algorithms (iterative deeds [7] and deep learning-based algorithm Convex Adam [24]). For each registration method 10 registered atlases per image are fed to the training routine expanding the T2-weighted training size from 40 to 400 label-image pairs each. Figure 4 shows a run with inline and out-of-line (see Sect. 2) data parameter training on the deeds registrations as an example how training scores behave.

Experiment IV: Consensus Generation and Subsequent Network Training. Using the training output of experiment III, we built 2 × 40 consensi: [10 deeds registered @ 40 fixed] and [10 Convex Adam registered @ 40 fixed]. Consensi were built by applying the STAPLE algorithm as baseline and opposed to that our proposed weighted-sum method on data parameters (DP) (see Sect. 2). On these, we trained several powerful nnUnet-models for segmentation [10]. In Fig. 5 in the foreground four box plots show the quality range of generated consensi regarding the oracle dice: [deeds, Convex Adam registrations]@[STAPLE, DP]. In the background the mean validation Dice of nnUnet-model trainings (150 epochs) is shown. As a reference, we trained directly on the T1-moving data with strong data augmentation (nnUNet "insane" augmentation) trying to overcome the domain gap directly (GAP). Furthermore, we trained on 40 randomly selected atlas labels (RND), all 400 atlas labels (ALL), STAPLE consensi, data parameter consensi (DP) and oracle-labels either on deeds or Convex Adam registered data. Note that the deeds data contained 40 unique moving atlases whereas the Convex Adam data contained 20 unique moving atlases, both warped to 40 fixed images as stated before (Fig. 3).

4 Results and Discussion

In **experiment I** we could show that our usage of data parameters is generally effective in the semantic segmentation scenario under artificial label noise. Figure 2 (middle) shows an increase of validation scores when activating stepping on data parameters after 10 epochs for disturbance strengths >0.1. Stronger disturbances lead to more severe score drops but can be recovered by using data parameters. In Fig. 2 (right) one can see that data parameters and oracle-Dice correlate most, when using the proposed risk regularization as well as the fixed weighting-scheme configuration (see Sect. 2). We did not notice any validation score improvements when switching between configurations and therefore conclude that a sorting of samples can also be learned inherently by the network. However, properly weighted data parameters can extract this information, make it explicitly visible and increase explainability. In **experiment II** we show that our approach works for registration noise during 2D training: When comparing different registration qualities, we observed that training scores drop from 81.1% to 69.6% Dice when lowering registration input quality. By using data parameters we can recover to a score of 74.1% meaning an improvement of +4.5%. **Experiment III** covers our target scenario—3D training with registered multi-atlas labels. With inline training of data parameters (used in the former experiments), validation scores during training drop significantly. Furthermore the data parameters do not separate high- and low quality registered atlases well (see Fig. 4, inline). When using our proposed out-of-line training approach (see Sect. 2) validation Dice and ranked correlation of data parameter values and oracle-Dice improve. **Experiment IV** shows that data parameters can be used to create a weighted-sum consensus as described in Sect. 2: Using data parameters, we can improve mean consensus-Dice for both, deeds and Convex Adam registrations over STAPLE [26] from 58.1% to 64.3% (+6.2%, ours, deeds data) and 56.8% to 61.6% (+4.8%, ours, Convex Adam data). When using the consensi in a subsequent nnUNet training [10], scores behave likewise (see Fig. 5). Regarding training times of over an hour with our LR-ASPP MobileNetV3-Large training, one has to consider that applying the STAPLE algorithm is magnitudes faster.

5 Conclusion and Outlook

Within this work, we showed that using data parameters in a multimodal prediction setting with propagated source labels is a valid approach to improve network training scores, get insight into training data quality and use the extracted info about sample quality in subsequent steps namely to generate consensus segmentations and provide these to further steps of deep learning pipelines. Our improvements over the original data parameter approach for semantic segmentation show strong results in both 2D- and 3D-training settings. Although we could extract sample quality information in the multi-atlas setting successfully, we could not improve network training scores in this setting directly since using the data parameters inline of the training loop resulted in unstable training. Regarding that, we want to continue investigating how an inline training can directly improve training scores in the multi-atlas setting. Furthermore our empirically chosen fixed weighting needs more theoretical foundation. The consensus generation could be further improved by trying more complex weighting schemes or incorporating the network predictions itself. Also we would like to compare our registration-segmentation pipeline against specialized approaches of Ding et al. and Liu et al. [4,5,15] which we consider as very interesting baselines.

References

1. Artaechevarria, X., Munoz-Barrutia, A., Ortiz-de Solorzano, C.: Combination strategies in multi-atlas image segmentation: application to brain MR data. IEEE Trans. Med. Imaging **28**(8), 1266–1277 (2009)
2. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: Proceedings of the 26th Annual International Conference on Machine Learning, pp. 41–48 (2009)
3. Castells, T., Weinzaepfel, P., Revaud, J.: SuperLoss: a generic loss for robust curriculum learning. Adv. Neural. Inf. Process. Syst. **33**, 4308–4319 (2020)
4. Ding, Z., Han, X., Niethammer, M.: VoteNet: a deep learning label fusion method for multi-atlas segmentation. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11766, pp. 202–210. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32248-9_23
5. Ding, Z., Han, X., Niethammer, M.: VoteNet+: an improved deep learning label fusion method for multi-atlas segmentation. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), pp. 363–367. IEEE (2020)
6. Heckemann, R.A., Hajnal, J.V., Aljabar, P., Rueckert, D., Hammers, A.: Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. Neuroimage **33**(1), 115–126 (2006)
7. Heinrich, M.P., Jenkinson, M., Brady, S.M., Schnabel, J.A.: Globally optimal deformable registration on a minimum spanning tree using dense displacement sampling. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) MICCAI 2012. LNCS, vol. 7512, pp. 115–122. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33454-2_15
8. Hempe, H., Yilmaz, E.B., Meyer, C., Heinrich, M.P.: Opportunistic CT screening for degenerative deformities and osteoporotic fractures with 3D DeepLab. In: Medical Imaging 2022: Image Processing. SPIE (2022)
9. Howard, A., et al.: Searching for mobilenetv3. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1314–1324 (2019)
10. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: NNU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)
11. Jiang, L., Zhou, Z., Leung, T., Li, L.J., Fei-Fei, L.: MentorNet: learning data-driven curriculum for very deep neural networks on corrupted labels. In: International Conference on Machine Learning, pp. 2304–2313. PMLR (2018)
12. Karimi, D., Dou, H., Warfield, S.K., Gholipour, A.: Deep learning with noisy labels: exploring techniques and remedies in medical image analysis. Med. Image Anal. **65**, 101759 (2020)

13. Kohl, S., et al.: A probabilistic U-Net for segmentation of ambiguous images. Adv. Neural Inf. Process. Syst. **31** (2018)
14. Liu, X., Song, L., Liu, S., Zhang, Y.: A review of deep-learning-based medical image segmentation methods. Sustainability **13**(3), 1224 (2021)
15. Liu, Z., et al.: Style curriculum learning for robust medical image segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12901, pp. 451–460. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87193-2_43
16. Loshchilov, I., Hutter, F.: SGDR: stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
17. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
18. 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, pp. 134–142 (2016)
19. Rohlfing, T., Russakoff, D.B., Maurer, C.R.: Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation. IEEE Trans. Med. Imaging **23**(8), 983–994 (2004)
20. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
21. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: MobileNetV2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520 (2018)
22. Saxena, S., Tuzel, O., DeCoste, D.: Data parameters: a new family of parameters for learning a differentiable curriculum. Adv. Neural Inf. Process. Syst. **32** (2019)
23. Shapey, J., et al.: Segmentation of vestibular schwannoma from magnetic resonance imaging: an open annotated dataset and baseline algorithm. The Cancer Imaging Archive (2021)
24. Siebert, H., Hansen, L., Heinrich, M.P.: Fast 3D registration with accurate optimisation and little learning for learn2Reg 2021. arXiv preprint arXiv:2112.03053 (2021)
25. Wang, H., Yushkevich, P.: Multi-atlas segmentation with joint label fusion and corrective learning-an open source implementation. Front. Neuroinform. **7**, 27 (2013)
26. Warfield, S.K., Zou, K.H., Wells, W.M.: Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. IEEE Trans. Med. Imaging **23**(7), 903–921 (2004)
27. Yan, W., et al.: The domain shift problem of medical image segmentation and vendor-adaptation by Unet-GAN. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 623–631. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_69
28. Zhang, Z., Zhang, H., Arik, S.O., Lee, H., Pfister, T.: Distilling effective supervision from severe label noise. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9294–9303 (2020)
29. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232 (2017)

A Method for Image Registration via Broken Geodesics

Alphin J. Thottupattu[1](✉), Jayanthi Sivaswamy[1], and Venkateswaran P. Krishnan[2]

[1] International Institute of Information Technology, Hyderabad 500032, India
alphinj.thottupattu@research.iiit.ac.in

[2] TIFR Centre for Applicable Mathematics, Bangalore 560065, India

Abstract. Anatomical variabilities seen in longitudinal data or inter-subject data is usually described by the underlying deformation, captured by non-rigid registration of these images. Stationary Velocity Field (SVF) based non-rigid registration algorithms are widely used for registration. However, these methods cover only a limited degree of deformations. We address this limitation and define an approximate metric space for the manifold of diffeomorphisms $\mathcal{G}$. We propose a method to break down the large deformation into finite set of small sequential deformations. This results in a broken geodesic path on $\mathcal{G}$ and its length now forms an approximate registration metric. We illustrate the method using a simple, intensity-based, log-demon implementation. Validation results of the proposed method show that it can capture large and complex deformations while producing qualitatively better results than state-of-the-art methods. The results also demonstrate that the proposed registration metric is a good indicator of the degree of deformation.

Keywords: Large deformation · Inter-subject registration · Approximate registration metric

1 Introduction

Computational anatomy is an area of research focused on developing computational models of biological organs to study the anatomical variabilities in the deformation space. Anatomical variations arise due to structural differences across individuals and changes due to growth or atrophy in an individual. These variations are studied using the deformation between the scans captured by a registration step. The registration algorithms typically optimize an energy functional based on a similarity function computed between the fixed and moving images.

Many initial image registration attempts use energy functionals inspired by physical processes to model the deformation as an elastic deformation [11], or

A. Hering et al. (Eds.): WBIR 2022, LNCS 13386, pp. 47–56, 2022.
https://doi.org/10.1007/978-3-031-11203-4_6

viscous flow [20] or diffusion [14]. The diffusion-based approaches have been explored for 3D medical images in general [6] and with deformations constrained to be diffeomorphic [23] to ensure preservation of the topology. The two main approaches used to capture diffeomorphisms are parametric and nonparametric methods. The Free Form Deformation (FFD) model [5,12] is a widely used parametric deformation model for medical image registration, where a rectangular grid with control points is used to model the deformation. Large diffeomorphic deformations [12] are handled by concatenating multiple FFDs. Deformable Registration via Attribute Matching and Mutual-Saliency Weighting (DRAMMS) [7] is a popular FFD-based method, which also handles inter-subject registration. DRAMMS matches Gabor features and prioritizes the reliable matching between images while performing registration. The main drawback of the deformations captured by FFD models is that they do not guarantee invertibility. The nonparametric methods represent the deformation with stationary or time varying velocity vector field. The diffeomorphic log-demon [23] is an example of the former while the Large Deformation Diffeomorphic Metric Mapping (LDDMM) [15] inspired from [8] is an example of the latter approach. In LDDMM, deformations are defined as geodesics on a Riemannian manifold, which is attractive; however, the methods based on this framework are computationally complex. The diffeomorphic log-demon framework [23], on the other hand, assigns a Lie group structure and assumes a stationary velocity field (SVF) which leads to computationally efficient methods, which is of interest to the community for practical purposes. This has motivated the exploration of a stationary LDDMM framework [16] that leverages the SVF advantage. The captured deformations are constrained to be symmetric in time-varying LDDMM [1] and log-demon [21] methods. Choosing an efficient optimization scheme such as Gauss-Newton as in [10] reduces the computational complexity of LDDMM framework. However, the log-demon framework is of interest to the community for practical purposes because of its computational efficiency and simplicity.

The Lie group structure gives a locally defined group exponential map to map the SVF to the deformation. Thus log-demon framework is meant to capture only neighboring elements in the manifold, i.e., only a limited degree of deformations can be captured. This will be referred to as the limited coverage issue of the SVF methods in this paper. Notwithstanding the limited coverage, several SVF based methods have been reported for efficient medical image registration with different similarity metrics, sim, such as local correlation between the images [17], spectral features [3], modality independent neighborhood descriptors [19] and wavelet features [9,18].

SVF based algorithms cannot handle complex deformations because the deformations are constrained to be smooth for the entire image and thus constrain the possible degree of deformation to some extent. We address this drawback by splitting the large deformation into finite set of smaller deformations. The key contributions of the paper are: i) an SVF-based registration framework to handle large deformations such as inter-subject variations computationally efficiently ii) an approximate metric to quantify structural variations between two images.

1.1 Background

Let G be a finite-dimensional Lie group with Lie algebra $\mathfrak{g}$. Recall that $\mathfrak{g}$ is the tangent space T_eG at the identity e of G. The exponential map $\exp : \mathfrak{g} \to G$ is defined as follows: Let $v \in \mathfrak{g}$. Then $\exp(v) = \gamma_v(1)$, where γ is the unique one-parameter subgroup of the Lie group G with v being its tangent vector at e. The vector v is called the infinitesimal generator of γ. The exponential map is a diffeomorphism from a small neighborhood containing 0 in the Lie algebra $\mathfrak{g}$ to a small neighborhood containing e of G.

Due to the fact that a bi-invariant metric may not exist for most of the Lie groups considered in medical image registration, the deformations considered here are elements of a Lie group with the Cartan-Schouten Connection [24]. This is the same as the one considered in the log-demon framework [23]. This is a left invariant connection [22] in which geodesics through the identity are one-parameter subgroups. The group geodesics are the geodesics of the connection. Any two neighboring points can be connected with a group geodesic. That is, if the stationary velocity field v connecting two images in the manifold $\mathcal{G}$ is small enough, then its group exponential map forms a geodesic. Similarly every $\mathfrak{g} \in G$ has a geodesically convex open neighbourhood [22].

2 Method

SVF based registration methods capture only a limited degree of deformation because exponential mappings are only locally defined. In order to perform registration of a moving image towards a fixed image, SVF is computed iteratively by updating it with a smoothed velocity field. This update is computed via a similarity metric that measures the correspondence between the moving and fixed images. The spatial smoothing has a detrimental effect as we explain next. A complex deformation typically consists of spatially independent deformations in a local neighbourhood. Depending on the smoothing parameter value, only major SVF updates in each region is considered for registration. Thus, modeling complex deformations with a smooth stationary velocity field is highly dependent on the similarity metric and the smoothing parameter in a registration algorithm. Finding an ideal similarity metric and an appropriate smoothing parameter applicable for any registration problem, irrespective of the complexity of the deformation and the type of data, is difficult.

We propose to address this issue as follows: Deform the moving image toward the fixed image by sequentially applying an SVF based registration. The SVF based algorithm chooses the major or the predominant (correspondence-based) deformation component among the spatially independent deformations in all the neighbourhoods to register along these predominant directions. The subsequent steps in the algorithm captures the next set of predominant directions sequentially. These sequentially captured deformations has a decreasing order of degree of pixel displacement caused by the deformations. Mathematically speaking, the discussion above can be summarised as follows. Consider complex deformations as a set of finite group geodesics and use a registration metric approximation to quantify the deformation between two images in terms of the length of a broken geodesic connecting them; a broken geodesic is a piecewise smooth curve, where each curve segment is a geodesic.

In the proposed method, the similarity-based metric selects the predominant deformation in each sequential step. The deformation that can bring the moving image in a step maximally closer to the target is selected from the one-parameter subgroup of deformations. In the manifold $\mathcal{G}$ every geodesic is contained in a unique maximal geodesic. Hence the maximal group geodesic γ_i computed using log-demon registration framework deforms the sequential image S_{i-1} in the previous step maximally closer to S_N. The maximal group geodesic paths are composed to get the broken geodesic path. As the deformation segments are diffeomorphic, the composed large deformation of the segments also preserves diffeomorphism to some extent.

In the proposed method, the coverage of the SVF method and the degree of deformation determines the number of subgroups N needed to cover the space. The feature based SVF methods in general, give more coverage for a single such subgroup and reduce the value of N.

A broken geodesic $\gamma : [0, T] \rightarrow M$ has finite number of geodesic segments γ_i for partitions of the domain $0 < t_1 < t_2 < \cdots < t_i < \cdots t_N = T$ where $i = 1, ...N$. The proposed algorithm to deform S_0 towards S_N is given in Algorithm 1. We have chosen the registration algorithm from [21] to compute SVF, u_i, in Algorithm 1. The Energy term is defined as: $\text{Energy}(S_i, S_N) = \text{sim}(S_N, S_i) + \text{Reg}(\gamma_i)$ where the first term is a functional of the similarity measure, which captures the correspondence between images, with $\text{sim}(S_N, S_i) = S_N - S_{i-1} \circ \exp(v_i)$. The second term is a regularization term, with $\text{Reg}(\gamma_i) = \|\bigtriangledown\gamma_i\|^2$.

Algorithm 1. Proposed Algorithm

```
Input: $S_0$ and $S_N$
Result: Transformation $\gamma = \exp(v_1) \circ \exp(v_2) \circ ... \exp(v_N)$
Initialization: $E_{\min}$=Energy$(S_0, S_N)$
repeat
    Register $S_{i-1}$ to $S_N \rightarrow u_i$
    Temp = $S_{i-1} \circ \exp(u_i)$
    $E_i$ = Energy(Temp, $S_N$)
    if $E_i < E_{\min}$ then
        $v_i = u_i$
        $E_{\min} = E_i$
        $S_i$ = Temp
    end if
until Convergence
```

2.1 Registration Metric Approximation

Let γ be a broken geodesic decomposed into N geodesics γ_i with stationary field v_i, i.e. $\dot{\gamma_i} = v_i(\gamma(t)) \in T_{\gamma_i(t)}M$. Each of the constant velocity paths γ_i is parameterized by the time interval $[t_{i-1}, t_i]$, and $N \in \mathbb{N}$ is minimized by requiring each of the geodesics in the broken geodesic to be maximal geodesics. The length of the broken geodesic is defined as,

$$l(\gamma) = \sum_{i}^{N} l(\gamma_i) = \sum_{i}^{N} d(S_{i-1}, S_i) \tag{1}$$

where, d is a distance metric defined in Eq. 2.

$$d(S_{i-1}, S_i) = \inf\{\|v_i\|_V, S_{i-1} \circ \exp(v_i) = S_i\}. \tag{2}$$

A registration metric needs to be defined to quantify the deformation between two images. The shape metric approximation in [25] can be used for the group geodesics of the Cartan-Schouten connection defined in the finite dimensional case as no bi-invariant metric exists. The length of a broken geodesic $l(\gamma)$ on the manifold $\mathcal{G}$ connecting S_0 and S_N, computed by Eq. 1 is defined as the proposed approximate metric.

3 Results

The proposed method was implemented using a simple intensity based log-demon technique [4] for illustrating the concept which is openly available at: http://dx.doi.org/10.17632/29ssbs4tzf.1. This choice also facilitates understanding the key strengths of the method independently. Two state-of-the-art (SOTA) methods are considered for performance comparison with the proposed method: the symmetric LDDMM implementation in ANTs [1] and DRAMMS which is a feature based, free-form deformation estimation method [7]. These two methods are considered to be good tools for inter-subject registration [26]. Publicly available codes were used for the SOTA methods with parameter settings as suggested in [26] for optimal performance. Both methods were implemented with B-spline interpolation, unless specified. 3D registration was done, and the images used in the experiments are 1.5T T1 MRI scans sourced from [2] and [13] unless specified otherwise. The number of maximum pieces in the broken geodesic path is set as five in all the experiments. The proposed image registration algorithm was used to register MRIs of different individuals.

3.1 Visual Assessment of Registration

To analyse the performance visually, six 3T MRI scans were collected. Three images collected from 20–30 year old male subjects were considered as moving images and three images collected from 40–50 year old female subjects were considered as fixed images. Performing a good registration is challenging with this selection of moving and fixed images. The high resolution MRI scans used for this experiment are openly available at http://dx.doi.org/10.17632/gnhg9n76nn.1. The registration results for these three different pairs are shown in Fig. 1-A. where only a sample slice is visualized for the 3 cases. The quality of registration can be assessed by observing the degree of match between images in the last two rows of each column. The mean squared error (MSE) was used as a similarity metric along with cubic interpolation. The results indicate that the proposed method is good at capturing complex inter-subject deformations.

The performance of the proposed method on medical images was compared with the state-of-the-art methods in Fig. 1-B. To apply the computed deformation, linear interpolation was used in all the methods. ANTs and the proposed method used MSE as a similarity metric for fair comparison and DRAMMS used its Gabor feature-based metric as it is a feature based method. The results shows that the deformations at the sulcal regions are better captured by the proposed method.

The quality of inverted deformations captured with ANTs and proposed method were also compared as follows. In Fig. 1-C the moving image deformed with moving-fixed deformation and fixed image deformed with inverted moving-fixed deformation are analysed for both the methods. The arrows overlaid on the registered images highlight regions where the proposed method yields error-free results as opposed to the other method. The results with proposed method shows better visual similarity with the target images in each case.

3.2 Quantitative Assessment of Registration

We present a quantitative comparison of the proposed method compared with ANTs and DRAMMS under the same setting. The average MSE for 10 image pair registrations with ANTs was 0.0036 ± 0.0009, with DRAMMS it was 0.0113 ± 0.0068 and with the proposed method it was $0.0012 \pm 7.0552e{-}08$.

The computed deformations in each method were used to transfer region segmentation (labels) from the moving image to the fixed image. The transferred segmentations are assessed using the Dice metric. Figure 2 shows a box plot of the obtained Dice values calculated by registering 10 pairs of brain MRIs with the fixed image, for white matter (WM), grey matter (GM) and 2 structures (L & R-hippocampus). The segmentation results for larger structures (i.e., WM and GM) are better with the proposed method compared to the other methods, whereas the smaller structure segmentation is comparable to DRAMMS.

3.3 Validation of Proposed Registration Metric

Finally, a validation of the proposed registration metric was done using two age-differentiated (20–30 versus 70–90 years) sets of MRIs, of 6 female subjects. Images from these 3D image sets were registered to an (independently drawn) MRI of a 20 year-old subject. The proposed registration metric was computed for the 6 pairs of registrations. A box plot of the registration metric value for each age group is shown in Fig. 3. Since the fixed image is that of a young subject,

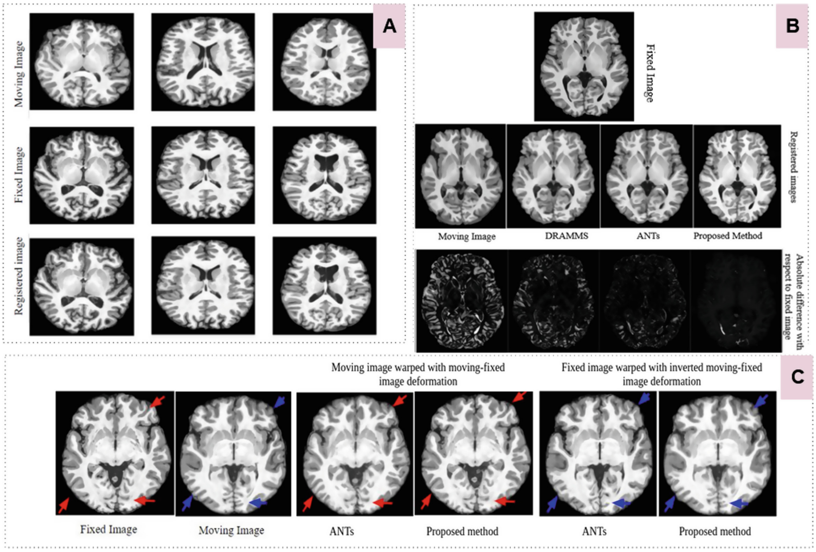


Fig. 1. A) Inter-subject image registration with proposed method for 3 pairs of volumes (in 3 columns) using cubic interpolation. Only sample slices are shown. B) Inter-subject image registration with 3 methods: DRAMMS, ANTs and the proposed method, implemented with linear interpolation. The regions near same colour arrows can be compared to check the registration accuracy. C) Forward and Backward Image Registration. Blue (Red) arrow shows where proposed method yields error-free results in moving (fixed) images, fixed (moving) images and warped moving (fixed) image using moving-fixed (inverted moving-fixed) deformation. Inverted moving-fixed deformation applied on fixed image and proposed method captures finer details compared to ANTs. (Color figure online)

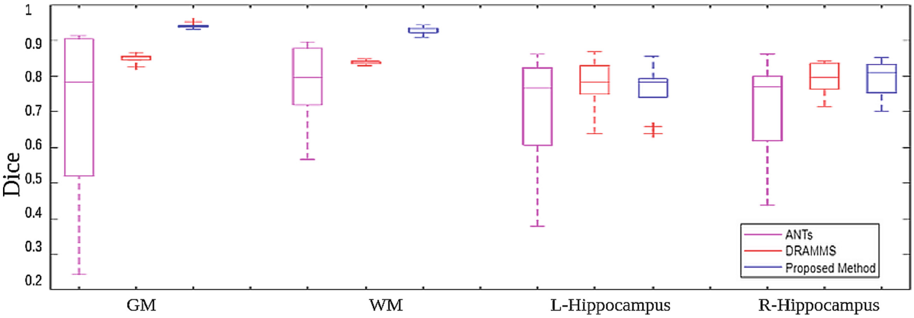


Fig. 2. Assessment of registration via segmentation of different structures using ANTS (magenta), DRAMMS (red), and the proposed method (blue). Box plots for the Dice coefficient are shown for White Matter (WM), Gray Matter (GM) and the Left and Right Hippocampi. (Color figure online)

the registration metric value should be higher for the older group than for the younger group, which is confirmed by the plot. Hence, it can be concluded that the proposed registration metric is a good indicator of natural deformations.

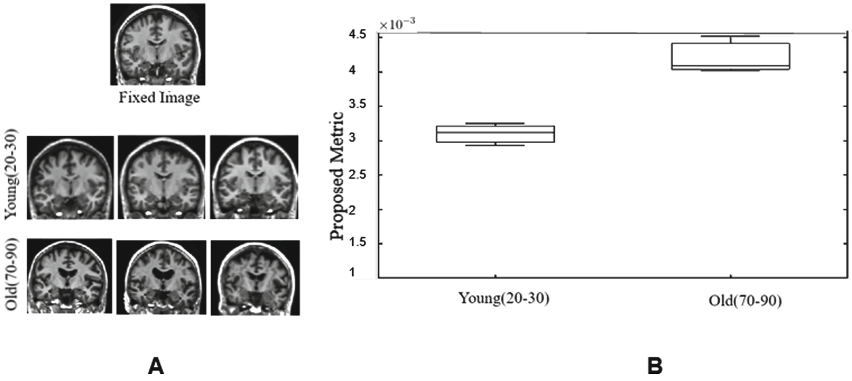


Fig. 3. Validation of the proposed registration metric. A) Central slices of images used to perform registration B) Box plots of the proposed registration metric values for registration of the fixed image with images of young and old subject group.

4 Discussion

Group exponential map based methods, with simple similarity registration metrics, fail to capture large deformations as the map is local in nature. We have addressed this issue in this paper by modelling large deformations with broken geodesic paths with the path length taken to be the associated registration metric. From the experiments it is observed that five pieces in the broken geodesic path is enough to capture very complex deformations. The proposed method does not guarantee diffeomorphism in a strict mathematical sense of infinite differentiability as the paths are modelled as piecewise geodesics. However, the experiments we have done suggest that the proposed method produces diffeomorphic paths. The results of implementation with a simple log-demon method show the performance to be superior to SOTA methods for complex/large deformations. We plan to extend this work by implementing the proposed framework using more efficient SVF based approaches such as in [3,9,17–19]. In summary, we have proposed a SVF-based registration framework that can capture large deformations and an approximate metric to quantify the shape variations between two images using the captured deformations.

References

1. Avants, B.B., et al.: Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Med. Image Anal. **12**, 26–41 (2008)

2. Landman, B., Warfield, S.: MICCAI 2012 Workshop on Multi-atlas Labeling, vol. 2. Create Space Independent Publishing Platform, Nice (2012)
3. Lombaert, H., Grady, L., Pennec, X., Ayache, N., Cheriet, F.: Spectral log-demons: diffeomorphic image registration with very large deformations. Int. J. Comput. Vision **107**(3), 254–271 (2013). https://doi.org/10.1007/s11263-013-0681-5
4. Lombaert, H.: Diffeomorphic Log Demons Image Registration. MATLAB Central File Exchange (2020)
5. Declerck, J., et al.: Automatic registration and alignment on a template of cardiac stress and rest reoriented SPECT images. IEEE Trans. Med. Imaging **16**, 727–37 (1997)
6. Pennec, X., Cachier, P., Ayache, N.: Understanding the "Demon's algorithm": 3D non-rigid registration by gradient descent. In: Taylor, C., Colchester, A. (eds.) MICCAI 1999. LNCS, vol. 1679, pp. 597–605. Springer, Heidelberg (1999). https://doi.org/10.1007/10704282_64
7. Ou, Y., et al.: DRAMMS: deformable registration via attribute matching and mutual-saliency weighting. Med. Image Anal. **15**(4), 622–639 (2011)
8. Trouvé, A.: Diffeomorphisms groups and pattern matching in image analysis. Int. J. Comput. Vision **28**, 213–221 (1998)
9. A, B: Good Afternoon. Conference, pp. 4–6 (2018)
10. Ashburner, J., Friston, K.J.: Diffeomorphic registration using geodesic shooting and gauss-newton optimisation. Neuroimage **55**, 954–967 (2011)
11. Broit, C.: Optimal registration of deformed images. University of Pennsylvania (1981)
12. Rueckert, D., et al.: Diffeomorphic registration using B-splines. Med. Image Comput. Comput. Assist. Interv. **9**(Pt 2), 702–709 (2006)
13. Hello and Goodbye: Good Evening. Journal 67 (2019)
14. Thirion, J.-P.: Image matching as a diffusion process: an analogy with Maxwell's demons. Med. Image Anal. **2**(3), 243–260 (1998)
15. Beg, M.F., et al.: Computing large deformation metric mappings via geodesic flows of diffeomorphisms. Int. J. Comput. Vision **61**, 139–157 (2005)
16. Hernandez, M., et al.: Registration of anatomical images using geodesic paths of diffeomorphisms parameterized with stationary vector fields. In: IEEE 11th International Conference on Computer Vision, pp. 1–8 (2007)
17. Lorenzi, M., et al.: LCC-Demons: a robust and accurate symmetric diffeomorphic registration algorithm. Neuroimage **81**, 470–483 (2013)
18. Pham, N., et al.: Spectral graph wavelet based nonrigid image registration. IEEE Trans. Pattern Anal. Mach. Intell., 3348–3352 (2018)
19. Reaungamornrat, S., et al.: MIND Demons: symmetric diffeomorphic deformable registration of MR and CT for image-guided spine surgery. IEEE Trans. Med. Imaging **35**(11), 2413–2424 (2016)
20. Christensen, T., et al.: Shoving model for viscous flow. World Sci. **12**, 375 (1981)
21. Vercauteren, T., et al.: Symmetric log-domain diffeomorphic Registration: a demons-based approach. Med. Image Comput. Comput. Assist. Interv. **11**(Pt 1), 754–761 (2008)
22. Arsigny, V., et al.: A fast and log-euclidean polyaffine framework for locally linear registration. [Research Report]RR-5885, INRIA (2006)
23. Arsigny, V., Commowick, O., Pennec, X., Ayache, N.: A log-euclidean framework for statistics on diffeomorphisms. In: Larsen, R., Nielsen, M., Sporring, J. (eds.) MICCAI 2006. LNCS, vol. 4190, pp. 924–931. Springer, Heidelberg (2006). https://doi.org/10.1007/11866565_113

24. Pennec, X.: Bi-invariant means on Lie groups with Cartan-Schouten connections. Geom. Sci. Inf., 59–67 (2013)
25. Yang, X., Li, Y., Reutens, D., Jiang, T.: Diffeomorphic metric landmark mapping using stationary velocity field parameterization. Int. J. Comput. Vision **115**(2), 69–86 (2015). https://doi.org/10.1007/s11263-015-0802-4
26. Ou, Y., et al.: Comparative evaluation of registration algorithms in different brain databases with varying difficulty: results and insights. IEEE Trans. Med. Imag. **33**, 2039–2065 (2014)

Deformable Image Registration Uncertainty Quantification Using Deep Learning for Dose Accumulation in Adaptive Proton Therapy

A. Smolders[1,2](✉), T. Lomax[1,2], D. C. Weber[1], and F. Albertini[1]

[1] Paul Scherrer Institute, Center for Proton Therapy, Villigen, Switzerland
andreas.smolders@psi.ch
[2] Department of Physics, ETH Zurich, Zürich, Switzerland

Abstract. Deformable image registration (DIR) is a key element in adaptive radiotherapy (AR) to include anatomical modifications in the adaptive planning. In AR, daily 3D images are acquired and DIR can be used for structure propagation and to deform the daily dose to a reference anatomy. Quantifying the uncertainty associated with DIR is essential. Here, a probabilistic unsupervised deep learning method is presented to predict the variance of a given deformable vector field (DVF). It is shown that the proposed method can predict the uncertainty associated with various conventional DIR algorithms for breathing deformation in the lung. In addition, we show that the uncertainty prediction is accurate also for DIR algorithms not used during the training. Finally, we demonstrate how the resulting DVFs can be used to estimate the dosimetric uncertainty arising from dose deformation.

Keywords: Deformable image registration · Proton therapy · Adaptive planning · Uncertainty · Deep learning

1 Introduction

Due to their peaked depth-dose profile, protons deposit a substantially lower dose to the normal tissue than photons for a given target dose [19]. However, the location of the dose peak is highly dependent on the tissue densities along the beam path, which are subject to anatomical changes throughout the treatment. Target margins are therefore applied, reducing the advantage of proton therapy (PT) [19]. The need to account for anatomical uncertainties can be alleviated using daily adaptive PT (DAPT), where treatment is reoptimized based on a daily patient image [1]. DAPT yields a series of dose maps, each specific to a daily anatomy. One important step of DAPT is to rely on the accurate accumulation of these doses for quality assurance (QA) of the delivered treatment and to trigger further adaptation [3,7,12,13]. To this end, the daily scans are registered to a reference and their corresponding doses are deformed before summation. In

A. Hering et al. (Eds.): WBIR 2022, LNCS 13386, pp. 57–66, 2022.
https://doi.org/10.1007/978-3-031-11203-4_7

the presence of deforming anatomy, deformable image registration (DIR) is used [7,22,25]. However, DIR is ill-posed [4], which results in dosimetric uncertainty after deformation. Substantial work has been performed to quantify this, summarized in [7], but there remains a clear need for methods predicting uncertainty associated with DIR and its effect on dose deformation [4,18].

In this work, an unsupervised deep learning (DL) method is presented to predict the uncertainty associated with a DIR result. Section 2 describes our method. The results of hyperparameter tuning on the predicted registration uncertainty are presented in Sect. 3, followed by the effect on the dosimetric uncertainty arising from dose deformation. Section 4 provides a discussion and conclusions are stated in Sect. 5.

2 Methods

Our work aims to estimate the uncertainty of the solution of an existing DIR algorithm. It is based upon a probabilistic unsupervised deep neural network for DIR called VoxelMorph [8]. The main equations from [8] are first summarized, after which the changes are described.

2.1 Probabilistic VoxelMorph

With f and m respectively a fixed and a moving 3D volume, here CT images, a neural network learns z, the latent variable for a parameterized representation of a deformable vector field (DVF) Φ_z. The network aims to estimate the conditional probability $p(z|f,m)$, by assuming a prior probability $p(z) = \mathcal{N}(0, \Sigma_z)$, with $\Sigma_z^{-1} = \Lambda_z = \lambda(D - A)$, λ a hyperparameter, D the graph degree matrix and A the adjacency matrix. Further, f is assumed to be a noisy observation of the warped moving image with noise level σ_I^2, $p(f|m,z) = \mathcal{N}(m \circ \Phi_z, \sigma_I^2 I)$. With these assumptions, calculation of $p(z|f,m)$ is intractable. Instead, $p(z|f,m)$ is modelled as a multivariate Gaussian

$$q_\Psi(z|f,m) = \mathcal{N}(\mu_{z|f,m}, \Sigma_{z|f,m}) \tag{1}$$

with Ψ the parameters of the network which predicts $\mu_{z|f,m}$ and $\Sigma_{z|f,m}$ (Fig. 1). The parameters Ψ are optimized by minimizing the KL divergence between $p(z|f,m)$ and $q_\Psi(z|f,m)$, yielding, for K samples $z_k \sim q_\Psi(z|f,m)$, a loss function

$$\begin{aligned}\mathcal{L}(\Psi, f, m) = \frac{1}{2\sigma_I^2 K} \sum_k ||f - m \circ \Phi_z||^2 + \frac{\lambda}{4} \sum_{i=1}^{m} \sum_{j \in N(i)} (\mu_i - \mu_j)^2 \\ + tr(\frac{\lambda}{2}(D - A)\Sigma_{z|f,m}) - \frac{1}{2} log(|\Sigma_{z|f,m}|) + cte\end{aligned} \tag{2}$$

with $N(i)$ the neighboring voxels of voxel i. When $\Sigma_{z|f,m}$ is diagonal, the last two terms of Eq. 2 reduce to $\frac{1}{2} tr(\lambda D \Sigma_{z|f,m} - log(\Sigma_{z|f,m}))$.

2.2 Combining Deep Learning with Existing DIR Software

Because the performance of DL based DIR is generally below conventional methods [9,10,24], our network aims to predict the uncertainty associated with a DVF generated by another algorithm without predicting the DVF itself. We therefore extend the VoxelMorph architecture to include the output DVF of an existing DIR algorithm (Fig. 1). First, an existing algorithm is ran on f and m, after which the resulting DVF is concatenated to f and m as network input. The network only predicts a diagonal matrix G, which is used to calculate $\Sigma_{z|f,m}$ (see Sect. 2.3), and the mean field $\mu_{z|f,m}$ is taken as the output of the DIR algorithm.

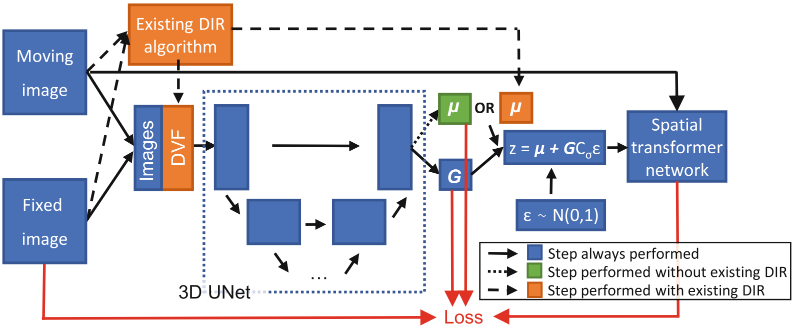


Fig. 1. Schematic network architecture. In case an existing DIR method is used, the resulting DVF of this algorithm is concatenated with the fixed and moving image, resulting in a 5 × H × W × D tensor as network input. A 3D UNet predicts a diagonal matrix G, and taking the DVF of the existing DIR as mean field μ, DVF samples are generated with the reparametrization trick as $z = \mu + GC_{\sigma_c}\epsilon$ (see Sect. 2.3) [15]. Contrarily if no existing DIR is used, the network only receives the fixed and moving image as input and predicts a mean DVF besides G.

2.3 Non-diagonal Covariance Matrix

Dosimetric uncertainty will be estimated by sampling $q_{\Psi}(z|f,m)$, requiring spatially smooth samples. Nearby vectors can be correlated with a non-diagonal covariance matrix. However, a full covariance matrix cannot be stored in memory because it would require storing $(3 \times H \times W \times D)^2$ entries, which for a 32 bit image of 256 × 265 × 96 requires 633 TB, compared to 25 MB for the diagonal elements. In [8] a non-diagonal $\Sigma_{z|f,m}$ is proposed by Gaussian smoothing of a diagonal matrix G, i.e. $\Sigma_{z|f,m} = C_{\sigma_c}GG^TC_{\sigma_c}^T$, but it is shown that this is unnecessary because the implemented diffeomorphic integration smooths the samples sufficiently. Because the existing DIR solutions are not necessarily diffeomorphic, we do not apply integration, which implies the need for a non-diagonal $\Sigma_{z|f,m}$.

Similar to [8], we apply Gaussian smoothing but invert the order $\Sigma_{z|f,m} = GC_{\sigma_c}C_{\sigma_c}^TG^T$ which yields a fixed correlation matrix $\rho = C_{\sigma_c}C_{\sigma_c}^T$. This has the advantage that the variance of the vector magnitude at voxel i is only dependent on the corresponding diagonal element $G_{i,i}$ and not on its neighbors. Furthermore, it allows to simplify the calculation of the loss terms in Eq. 2. Rewriting the last two terms of Eq. 2 with $\Sigma_{z|f,m} = GC_{\sigma_c}C_{\sigma_c}^T G^T$ results in

$$\begin{aligned} & tr(\frac{\lambda}{2}(D-A)\Sigma_{z|f,m}) - \frac{1}{2}log(|\Sigma_{z|f,m}|) \\ & = \sum_{i=1}^{m}\sum_{j=1}^{m}(\frac{\lambda}{2}(D-A)_{i,j}(\Sigma_{z|f,m})_{i,j}) - \frac{1}{2}log(|GG^T|) - \frac{1}{2}log(|C_{\sigma_c}C_{\sigma_c}^T|) \end{aligned} \quad (3)$$

with i and j respectively the row and column indices, m the number voxels and $log(|C_{\sigma_c}C_{\sigma_c}^T|)$ a constant which can be excluded from the loss function. For each row (or voxel) i, the matrix $(D-A)$ has only 7 non-zero elements (the voxel itself and its 6 neighboring voxels), so that only the corresponding 7 elements in $\Sigma_{z|f,m}$ are needed to evaluate the loss function. By precomputing the 7 corresponding elements of $\rho = C_{\sigma_c}C_{\sigma_c}^T$, the first term of Eq. 3 becomes

$$\sum_{i=1}^{m}\sum_{j=1}^{m}(\frac{\lambda}{2}(D-A)_{i,j}(\Sigma_{z|f,m})_{i,j}) = \sum_{i=1}^{m}\sum_{j\in N(i)}(\frac{\lambda}{2}(D-A)_{i,j}\rho_{i,j}G_{i,i}G_{j,j}) \quad (4)$$

with $N(i)$ the neighbors of voxel i, which allows fast evaluation of $\mathcal{L}$ without the need of storing large matrices.

2.4 Training

52 CT scan pairs from 40 different patients with various indications treated at the Centre for Proton Therapy (CPT) in Switzerland are used for training. The pairs consist of one planning and one replanning or control CT from a proton treatment, and are therefore representative of both daily and progressive anatomical variations in DAPT. Scans are rigidly registered using the `Elastix` toolbox [16] and resampled to a fixed resolution 1.95 × 1.95 × 2.00 mm, most frequently occurring in the dataset. The Hounsfield units are normalized with $\frac{HU+1000}{4000}$. Patches with a fixed size $256 \times 256 \times 96$ are randomly cropped from the full CTs during training and axis aligned flipping is applied as data augmentation.

The network is implemented in `Pytorch` [20] and training is ran on GPUs with 11 GB VRAM. A 3D UNet is used [8] with an initial convolution creating 16 feature maps, which are doubled in each of the 3 consecutive downsampling steps. The features are upsampled 3 times to their original resolution. The parameters are optimized with `Adam` [14] with initial learning rate $2 \cdot 10^{-4}$, which is halved 6 times during 500 epochs. Gaussian smoothing of the diagonal covariance matrix has a fixed kernel size of 61 voxels and blur $\sigma_c = 15$.

We train networks to predict the uncertainty associated with three existing DIR algorithms: a b-spline and a demon implementation in Plastimatch and a non-diffeomorphic VoxelMorph predicting both $\mu_{z|f,m}$ and $\Sigma_{z|f,m}$. The parameters for b-spline and demon are taken from [2,17]. Furthermore, we verify whether these networks can be used to predict the uncertainty of other DIR algorithms by evaluating them on the results of a commercial DIR in Velocity.

2.5 Validation

The hyperparameters λ and σ_I^2 are tuned for each method by quantitatively evaluating the predicted uncertainty on the publicly available 4DCT DIRLAB lung deformation dataset [5,6]. It contains 10 CT scan pairs with each 300 annotated landmarks (LM). These scans are split equally in a validation and test set. We maximize the probability of observing the moving landmarks $\boldsymbol{x}_m$ given the predicted probabilistic vector field, which, for a given set of CTs, is calculated as

$$p(LMs) = \prod_{i}^{CTs} \prod_{j}^{LM} p(\boldsymbol{x}_{m,i,j}|DVF_i), \tag{5}$$

assuming for simplicity that each landmark is independent of the others, which is reasonable if the landmarks are sufficiently far apart. Note that the probability of observing exactly $\boldsymbol{x}_m$ is infinitesimally small because the variables are continuous. We therefore maximize the probability that $\boldsymbol{x}_m$ is observed within a cube of 1 mm^3 around it with a homogeneous probability density, which is the same as maximizing the probability density at $\boldsymbol{x}_m$. We discard the 1% least probable points because $p(LMs)$ is heavily affected by the outliers due to the extremely low probability density at the tails of a normal distribution. Furthermore, we maximize the mean log $p(LMs)$ to avoid that the absolute value is dependent on the number of landmarks.

3 Results

3.1 Hyper Parameter Tuning

The optimal hyperparameters are $\lambda = 10$ and $\sigma_I^2 = 10^{-4}$ for both b-spline and demon (Fig. 2). Further, using both the networks trained on demon and b-spline, we find that the network trained with b-spline and $\lambda = 5$ and $\sigma_I^2 = 10^{-4}$ yields the highest average log $p(LMs)$ for Velocity (not shown).

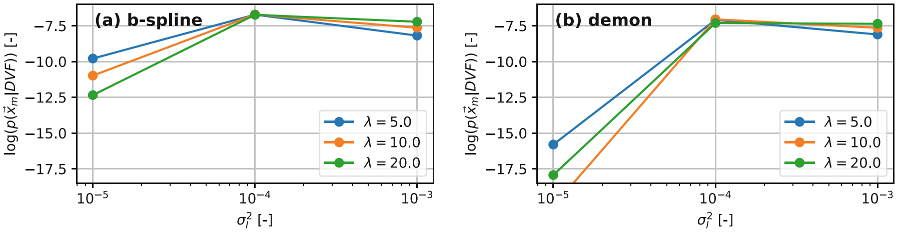


Fig. 2. Average log probability of observing moving landmarks $\boldsymbol{x}_m$ of the validation set for varying values of σ_I^2 and λ including an existing DIR output. Similar results were found for the test set (not shown).

For VoxelMorph, the hyperparameters influence both $\mu_{z|f,m}$ and $\Sigma_{z|f,m}$. Equation 2 shows that the trade off between similarity and smoothness is determined by the product $\lambda\sigma_I^2$. Therefore, we first minimize the target registration error (TRE) on the validation set by varying $\lambda\sigma_I^2$ (keeping $\lambda = 2$), which yields a minimum TRE around $\lambda\sigma_I^2 = 2 \cdot 10^{-3}$. Varying λ and σ_I^2 while keeping $\lambda\sigma_I^2 = 2{\cdot}10^{-3}$ results in a maximum $p(LMs)$ for $\sigma_I^2 = 5{\cdot}10^{-4}$. $p(LMs)$ is however lower than for the networks including the conventional (i.e. non deep learning) DIRs, indicating that these methods predict better probability distributions.

Figure 3 shows the uncertainty prediction for a lung CT in the DIRLAB dataset. As expected, the predicted uncertainty is low in regions with high contrast and high where contrast is low. Further, the Jacobian determinant is <0 for on average 0.01% of the voxels in sampled DVFs for the DIRLAB dataset, which, together with visual inspection, indicates that samples are sufficiently smooth.

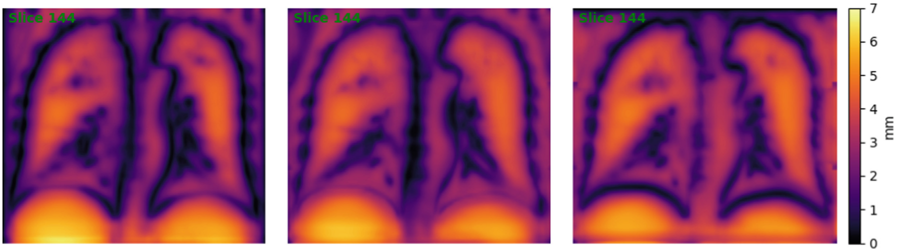


Fig. 3. Predicted uncertainty σ_p, i.e. the square root of the diagonal elements of $\Sigma_{z|f,m}$, in the sagittal (left), coronal (middle) and axial (right) direction for one example patient in the test set.

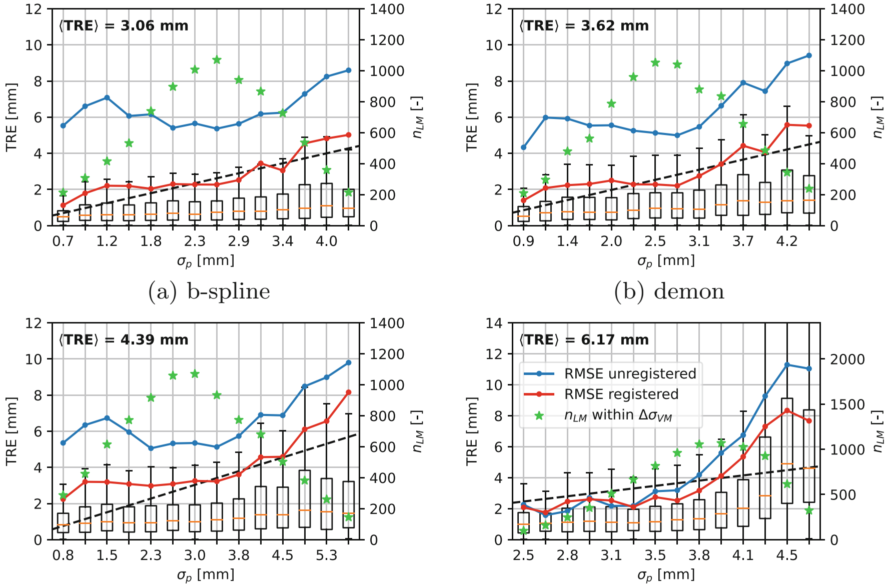
Comparing the target errors and their predictions for the tuned networks for all DIRLAB scans yields several conclusions (Fig. 4). First of all, our method is able to fairly accurately predict the uncertainty of multiple existing DIR algorithms. Secondly, the error prediction of Velocity shows that it is possible to predict the error from a DIR algorithm even if it was not used to train the network. Lastly, the average error is higher and the uncertainty prediction is worse for VoxelMorph than for the existing DIR algorithms, as expected from [9,10,24]. However, the performance can likely be improved by diffeomorphic integration, network adjustments or using more data, but this is not within the scope of the current study.

3.2 Dose Deformation

We create probabilistic dose maps by sampling the probabilistic DVF and warping the dose with the different samples. We focus here on the result of a single deformation to highlight the dosimetric uncertainty associated with warping. Even though the predicted DVFs have assumed to be Gaussian, the probabilistic dose maps are not. We therefore keep the individual samples and use a finite-sample distribution to approximate the probabilistic dose map.

The dose received by the tumor and organs at risk (OARs) is in PT frequently evaluated with dose volume histograms (DVHs). Probabilistic DVHs can be constructed from the probabilistic dose map. Here, the lower and upper bound of the DVH depict for each volume increment respectively the 5th and 95th percentile of all sampled doses (Fig. 5).

Verifying whether the dosimetric uncertainty is realistic is non-trivial. Previous work [2,17] quantified it by warping the dose with several DIR algorithms and calculating the dose differences between the results. Similarly, here we verify whether the warped dose with three conventional DIR algorithms falls in



(a) b-spline

(b) demon

(c) velocity

(d) VoxelMorph

Fig. 4. TRE as a function of the predicted uncertainty σ_p for all DIRLAB scans. For each subplot, σ_p is divided into 15 equal intervals and the distribution of the TREs within each interval is plotted as a box, together with the unregistered and registered root mean squared error (RMSE). The number of landmarks n_{LM} within each interval is also shown (right axes). If the TREs were normally distributed and the networks had a perfect prediction, the registered RMSE would be exactly equal to the predicted σ_p (dashed line).

between our predicted lower and upper bound (Fig. 5). Using the same dataset of 7 lung cancer patients with each 9 repeated CTs as in [2,17], we find that the dose in on average 97% of the volume of the OARs (heart, esophagus and medulla) lies between the bounds predicted for b-spline. For the planning target volume (PTV) and gross tumor volume (GTV) it is on average 81%.

4 Discussion

Despite the promising preliminary results, more work is required before the method can be used in the clinic. Our approach should be verified on a dataset including typical deformations that occur during the course of six weeks of treatment, and not only during one breathing cycle. To that end, a dataset with typical anatomical deformations is currently being landmarked at the CPT.

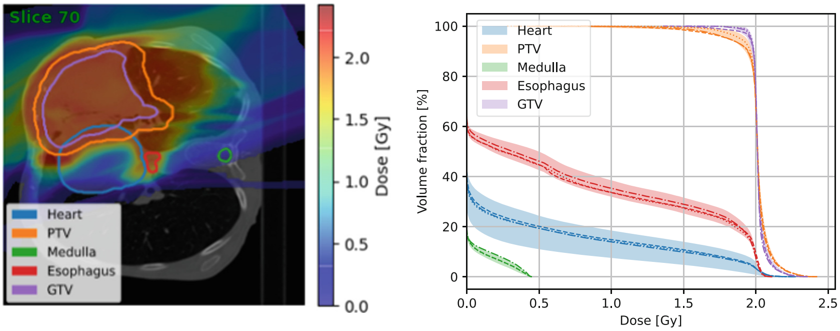


Fig. 5. Left: example of a deformed dose map with b-spline, overlayed with contours of the gross tumor volume (GTV), planning target volume (PTV) and three OARs. Right: corresponding probabilistic DVH as calculated with the optimal network for b-spline (shaded area). The dashed, dotted and dash-dotted lines represent the DVH for warped doses with three commercial DIR softwares, respectively Mirada, Raystation Anaconda and Velocity.

Even for the dataset under study, the error prediction is clearly not perfect. This can be due to several factors, among which imperfect annotation, lack of training data or inaccurate model assumptions. One important assumption is the Gaussian vector field. Although our results show that it is not unreasonable to assume that the errors are Gaussian, further research should look whether other probability distributions yield better results. Unfortunately, other analytical distributions are often mathematically more complex making exact treatment as in Eqs. 2 and 3 difficult. Learning a discretized posterior could resolve this [10,11,21,23].

The trained networks capture most of the dosimetric variations found in the OARs when running conventional DIRs. By contrast, for the GTV and PTV only 81% of the doses lie between the error bars, significantly below the expected

90% given the 5th and 95th percentile error bounds. However, we found that this value increases to 91% by simply adding a small margin to the error bounds (i.e. by increasing the upper and decreasing lower bound by only 0.1% of the dose). This indicates that the deviation from the error bounds is mostly very small.

5 Conclusion

In this work, a probabilistic unsupervised deep learning method for deformable image registration is presented to predict the uncertainty associated with DIR solutions. It is shown that the method can accurately predict the uncertainty of various conventional DIR algorithms and that the combination of deep learning with conventional DIR yields superior results than using deep learning alone.

Acknowledgments. This work has received funding from the European Union's Horizon 2020 Marie Skłodowska-Curie Actions under Grant Agreement No. 955956.

References

1. Albertini, F., Matter, M., Nenoff, L., Zhang, Y., Lomax, A.: Online daily adaptive proton therapy. Br. J. Radiol. **93**(1107), 20190594 (2020)
2. Amstutz, F., et al.: An approach for estimating dosimetric uncertainties in deformable dose accumulation in pencil beam scanning proton therapy for lung cancer. Phys. Med. Biol. **66**(10), 105007 (2021)
3. Brock, K.K., McShan, D.L., Ten Haken, R., Hollister, S., Dawson, L., Balter, J.: Inclusion of organ deformation in dose calculations. Med. Phys. **30**(3), 290–295 (2003)
4. Brock, K.K., Mutic, S., McNutt, T.R., Li, H., Kessler, M.L.: Use of image registration and fusion algorithms and techniques in radiotherapy: report of the AAPM radiation therapy committee task group no. 132. Med. Phys. **44**(7), e43–e76 (2017)
5. Castillo, E., Castillo, R., Martinez, J., Shenoy, M., Guerrero, T.: Four-dimensional deformable image registration using trajectory modeling. Phys. Med. Biol. **55**(1), 305 (2009)
6. Castillo, R., et al.: A framework for evaluation of deformable image registration spatial accuracy using large landmark point sets. Phys. Med. Biol. **54**(7), 1849 (2009)
7. Chetty, I.J., Rosu-Bubulac, M.: Deformable registration for dose accumulation. In: Seminars in Radiation Oncology, vol. 29, pp. 198–208. Elsevier (2019)
8. Dalca, A.V., Balakrishnan, G., Guttag, J., Sabuncu, M.R.: Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. Med. Image Anal. **57**, 226–236 (2019)
9. Hansen, L., Heinrich, M.P.: Tackling the problem of large deformations in deep learning based medical image registration using displacement embeddings. arXiv preprint arXiv:2005.13338 (2020)
10. 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

11. Heinrich, M.P., Jenkinson, M., Papież, B.W., Brady, S.M., Schnabel, J.A.: Towards realtime multimodal fusion for image-guided interventions using self-similarities. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013. LNCS, vol. 8149, pp. 187–194. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40811-3_24
12. Jaffray, D.A., Lindsay, P.E., Brock, K.K., Deasy, J.O., Tomé, W.A.: Accurate accumulation of dose for improved understanding of radiation effects in normal tissue. Int. J. Radiation Oncol.* Biol.* Phys. **76**(3), S135–S139 (2010)
13. Janssens, G., et al.: Evaluation of nonrigid registration models for interfraction dose accumulation in radiotherapy. Med. Phys. **36**(9Part1), 4268–4276 (2009)
14. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
15. Kingma, D.P., Salimans, T., Welling, M.: Variational dropout and the local reparameterization trick. Adv. Neural Inf. Process. Syst. **28** (2015)
16. Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P.: Elastix: a toolbox for intensity-based medical image registration. IEEE Trans. Med. Imaging **29**(1), 196–205 (2009)
17. Nenoff, L., et al.: Deformable image registration uncertainty for inter-fractional dose accumulation of lung cancer proton therapy. Radiother. Oncol. **147**, 178–185 (2020)
18. Paganelli, C., Meschini, G., Molinelli, S., Riboldi, M., Baroni, G.: Patient-specific validation of deformable image registration in radiation therapy: overview and caveats. Med. Phys. **45**(10), e908–e922 (2018)
19. Paganetti, H.: Range uncertainties in proton therapy and the role of Monte Carlo simulations. Phys. Med. Biol. **57**(11), R99 (2012)
20. Paszke, A., et al.: PyTorch: an imperative style, high-performance deep learning library. Adv. Neural. Inf. Process. Syst. **32**, 8026–8037 (2019)
21. Rühaak, J.: Estimation of large motion in lung CT by integrating regularized keypoint correspondences into dense deformable registration. IEEE Trans. Med. Imaging **36**(8), 1746–1757 (2017)
22. Schultheiss, T.E., Tomé, W.A., Orton, C.G.: It is not appropriate to "deform" dose along with deformable image registration in adaptive radiotherapy. Med. Phys. **39**(11), 6531–6533 (2012)
23. Sedghi, A., Kapur, T., Luo, J., Mousavi, P., Wells, W.M.: Probabilistic image registration via deep multi-class classification: characterizing uncertainty. In: Greenspan, H., et al. (eds.) CLIP/UNSURE 2019. LNCS, vol. 11840, pp. 12–22. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32689-0_2
24. 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. Med. Image Anal. **52**, 128–143 (2019)
25. Zhong, H., Chetty, I.J.: Caution must be exercised when performing deformable dose accumulation for tumors undergoing mass changes during fractionated radiation therapy. Int. J. Radiat. Oncol. Biol. Phys. **97**(1), 182–183 (2016)

Distinct Structural Patterns of the Human Brain: A Caveat for Registration

Frithjof Kruggel(✉)

Department of Biomedical Engineering, University of California, Irvine, USA
fkruggel@uci.edu
http://sip.eng.uci.edu

Abstract. Current approaches for analyzing structural patterns of the human brain often implicitly assume that brains are variants of a single type, and use nonlinear registration to reduce the inter-individual variability. This assumption is challenged here. Regional anatomical and connection patterns cluster into statistically distinct types. An advanced analysis proposed here leads to a deeper understanding of the governing principles of cortical variability.

Keywords: Structural patterns · Connectivity · Human cortex

1 Introduction

Cortical structures of the human brain show a puzzling complexity and inter-individual variability. Numerous analytic approaches implicitly assume that structural properties of brains, represented in any high-dimensional space, form a single cluster and use nonlinear registration to reduce the inter-individual variability. We challenge this assumption. Depending on the features and similarity criteria involved in the registration process, the total variance is reduced by only 20–40%. Consider a simplifying analogy: Suppose we want to study structural properties of cars. We hardly doubt that a registration procedure can be designed that successfully matches gross car parts (e.g., the passenger and engine compartment, the trunk and wheels). However, when zooming into details, objects under study become distinct (e.g. a trunk of a truck vs. a sports car, a combustion engine vs. an electric motor). Here, we demonstrate here that structural variants of brain regions with distinctive properties exist in a population. Avoiding an arguable registration and embracing the actual variability leads to analytic procedures that actually *explain* sources of variability at a considerably larger proportion.

2 Methods

Data Source: We used anatomical and diffusion-weighted MRI data acquired in $nc = 1061$ subjects of the publicly available Human Connectome Project [2].

A. Hering et al. (Eds.): WBIR 2022, LNCS 13386, pp. 67–71, 2022.
https://doi.org/10.1007/978-3-031-11203-4_8

Anatomical processing: We started out from triangulated meshes representing the white-gray matter interface of a hemisphere with a topological genus of zero. Using local curvature and geodesic depth, the surface was segmented into patches called *basins* that were centered around a locally deepest point, the *sulcal root.* A most isometric mapping was used to transfer and re-parameterize vertex-wise properties (e.g., basin label, depth, curvature) onto a common sphere with $nv = 163842$ vertices. Thus, we represented structural information as an image of $nc \times nv \times np$ properties. Refer to [4] for details.

Tractography: Diffusion-weighted data were corrected for subject motion and susceptibility distortions. Voxel-wise estimates of the orientation distribution function of water mobility were computed using the constrained spherical deconvolution method [3]. Probabilistic tracking [5] from basin-labeled surface seeds was performed to determine connectivity between basins. Results were kept in hemisphere-wise connectivity matrices C, where each element $C(i, j)$ corresponded to the probability of connecting basin i to j. Thus, C can be regarded as a discrete, empirical PDF of basin connectivity.

Distance Metrics: We computed a co-occurence matrix M of the basin labeling in hemispheres a, b and expressed the their structural distance by $D_M = 1 - \mathrm{NMI}(M_{a,b})$. For connectivity, we selected the Hellinger distance metric by experimentation:

$$\mathrm{D_C}(a, b) = \sqrt{2 \left(1 - \sqrt{2 \sum_i \sum_j \sqrt{C_a(i, j)\, C_b(i, j)}} \right)}. \tag{1}$$

Statistical Assessment: We computed the distance metrics for all hemisphere pairs a, b and compiled them in matrices for structure D_M and connectivity D_C of dimensions $nc \times nc$. Both matrices were mapped into a low-dimensional space using the ISOMAP algorithm [6], with a target dimension of $nd = 4$ estimated by the Grassberger-Procaccia method [1]. Thus, structural and connectivity of a hemisphere were represented by a point in an 8-dimensional space. We used a Gaussian mixture model to cluster into groups, where the number of classes was determined from the maximal Bayesian information criterion and silhouette coefficient. Note that this analysis can be restricted to any sub-region of the whole hemisphere.

3 Results

Due to space limitations, we provided results for the central sulcus (CS) only. For each dimension of the structural and connectivity matrices, we analyzed their dependence on several variables using linear regression (Tab. 1). Dimensions and their amount of represented variance were compiled in the second column. The first dimension represented more than 50% of the variance, and corresponded to the "regularity" of the sulcus structure. Regular sulci were straight, deep, and consisted of relatively few basins, in contrast to tortuous, shallow sulci with a larger number of basins (Fig. 1). Considering the number of basins as a proxy for structural regularity, we found that between 25% and 41% of the variance (R^2) were addressed to regularity. About 10% of the overall variance were explained by subject sex, handedness, and brain volume.

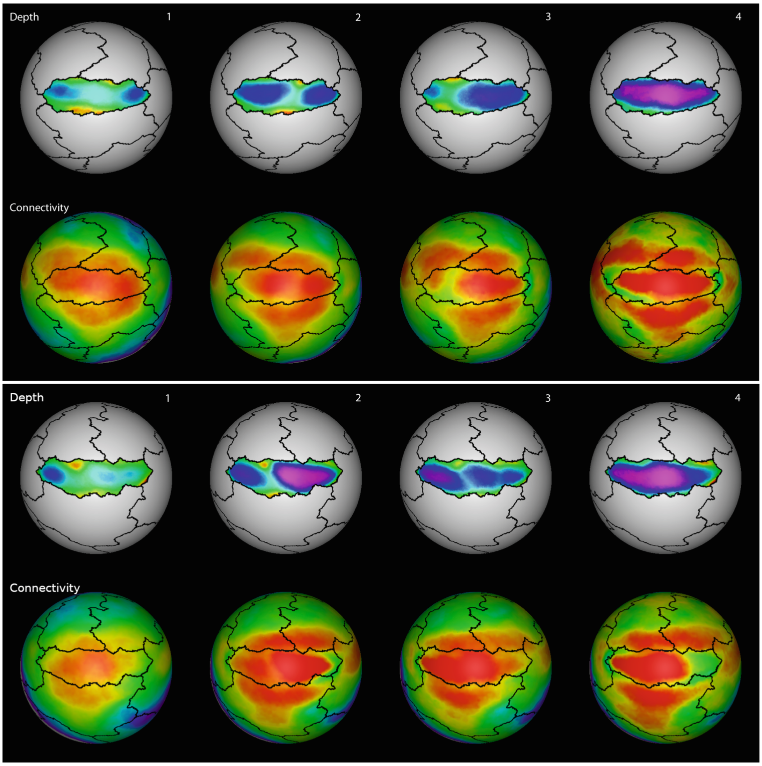


Fig. 1. Clustering of the central sulcus (CS) into four distinct, mirror-symmetric configurations on the left (top) and right (below) side. Rows 1, 3 show geodesic depth (increasing from red to magenta). Rows 2, 4 show the connection strength (increasing from magenta to red). (Color figure online)

Significant influences of subject sex, handedness, and brain volume were typically found for the second structural dimension and the third connectivity dimension. We assessed the absolute difference of scores within subject pairs grouped by genetic similarity. This heritability was typically reflected in the second dimension, representing between 2% and 6% of the total variance.

Clustering yielded four distinct structural and connectivity patterns (Fig. 1), with mirror-symmetric patterns on the left (top panel) and right side (below). Patterns were sorted by increasing regularity from left to right, as determined

Table 1. Analysis of dimensions obtained from domain decomposition of distance matrices for the central sulcus on the left and right side. The relevance of dimensions 1–4 was assessed in relation to the number of basins in this sulcus, demographic variables sex, handedness, and heritability.

Model	Dimension		# of Basins			Brain Volume		Sex		Handedness		Heritability	
		Exp. var.	p-value	Code	R^2	p-value	Code	p-value	Code	p-value	Code	p-value	Code
Structure left	1	0.559	< 2e–16	***	0.337	n.s	–	n.s	–	n.s	–	n.s	–
	2	0.123	8.13e–4	***	0.026	1.95e–5	***	n.s	–	n.s	–	1.05e–4	***
	3	0.084	4.29e–6	***	0.037	5.51e–3	**	1.54e–3	**	0.0149	*	n.s	–
	4	0.045	< 2e–16	***	0.127	n.s	–	n.s	–	n.s	–	n.s	–
Structure right	1	0.588	< 2e–16	***	0.255	n.s	–	n.s	–	0.0413	*	n.s	–
	2	0.098	0.0170	*	0.038	2.48e–8	***	1.13e–4	**	n.s	–	0.0349	*
	3	0.058	1.49e–13	***	0.054	0.016	*	n.s	–	n.s	–	n.s	–
	4	0.046	< 2e–16	***	0.194	n.s	–	n.s	–	n.s	–	n.s	–
Connectivity left	1	0.562	< 2e–16	***	0.406	n.s	–	7.40e-3	**	n.s	–	n.s	–
	2	0.230	< 2e–16	***	0.139	n.s	–	n.s	–	n.s	–	4.44e-3	**
	3	0.106	0.0378	*	0.020	2.69e–3	**	8.35e–3	**	6.29e–3	**	n.s	–
	4	0.030	n.s	–	0.024	5.45e-5	***	n.s	–	n.s	–	0.0451	*
Connectivity right	1	0.528	< 2e–16	***	0.380	n.s	–	n.s	–	n.s	–	n.s	–
	2	0.253	< 2e–16	***	0.111	n.s	–	n.s	–	n.s	–	n.s	–
	3	0.119	1.52e–5	***	0.025	n.s	–	0.0122	*	n.s	–	n.s	–

from scores of the first dimension above. The first pattern (column 1) showed a low regularity, consisting of two shallow centers of low variability. Patterns 2 and 3 revealed two stronger centers, in pattern 2 more prominent in the upper CS, in pattern 3 more prominent in the lower CS. Finally, pattern 4 showed a straight and deep sulcus with high regularity. Interestingly, more regular sulcal patterns were related to a stronger, more distinctive connectivity (rows 2 and 4). Note that connection strength closely followed a lower basin variability not only in the central sulcus, but also adjacent regions in the pre- and post-central sulcus, and the mid-posterior insula on both sides.

4 Conclusion

By this short demonstration, we wanted to illustrate two points: (1) Structural and connectivity patterns of the human brain do not originate from a continuum, but show distinct properties, at least at the regional level. This finding renders the application of registration processes as arguable, at least at the hemispheric level. (2) Instead of attempting to reduce the inter-individual variability by registration, we suggest to embrace this variability and to analyze and identify their sources. As demonstrated here, up to 80% of the total variance can be explained by identifiable factors.

References

1. Grassberger, P., Procaccia, I.: Characterization of strange attractors. Physica D: Nonlinear Phenomena **9**, 189–208 (1983)
2. Human Connectome Project: 1200 Subjects Data Release Reference Manual. https://www.humanconnectome.org/study/hcp-young-adult/document/1200-subjects-data-release, Accessed 17 Apr 2022
3. Jeurissen, B., Tournier, J.D., Dhollander, T., Connelly, A., Sijbers, J.: Multi-tissue constrained spherical deconvolution for improved analysis of multi-shell diffusion MRI data. NeuroImage **103**, 411–426 (2014)
4. Kruggel, F.: The macro-structural variability of the human neocortex. NeuroImage **172**, 620–630 (2018)
5. Smith, R.E., Tournier, J.D., Calamante, F., Connelly, A.: Anatomically-constrained tractography: improved diffusion MRI streamlines tractography through effective use of anatomical information. NeuroImage **62**, 1924–1938 (2016)
6. Tenenbaum, J.B., de Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science **290**, 2319–2323 (2000)

Architectures

A Multi-organ Point Cloud Registration Algorithm for Abdominal CT Registration

Samuel Joutard[1,3](✉), Thomas Pheiffer[3], Chloe Audigier[2], Patrick Wohlfahrt[2], Reuben Dorent[1], Sebastien Piat[3], Tom Vercauteren[1], Marc Modat[1], and Tommaso Mansi[3]

[1] King's College London, London, UK
samuel.joutard@kcl.ac.uk
[2] Siemens Healthineers, Erlangen, Germany
[3] Siemens Healthineers, Princeton, USA

Abstract. Registering CT images of the chest is a crucial step for several tasks such as disease progression tracking or surgical planning. It is also a challenging step because of the heterogeneous content of the human abdomen which implies complex deformations. In this work, we focus on accurately registering a subset of organs of interest. We register organ surface point clouds, as may typically be extracted from an automatic segmentation pipeline, by expanding the Bayesian Coherent Point Drift algorithm (BCPD). We introduce MO-BCPD, a multi-organ version of the BCPD algorithm which explicitly models three important aspects of this task: organ individual elastic properties, inter-organ motion coherence and segmentation inaccuracy. This model also provides an interpolation framework to estimate the deformation of the entire volume. We demonstrate the efficiency of our method by registering different patients from the LITS challenge dataset. The target registration error on anatomical landmarks is almost twice as small for MO-BCPD compared to standard BCPD while imposing the same constraints on individual organs deformation.

1 Introduction

Registering CT images of the chest is an important step for several pipelines such as surgical planning for liver cancer resection or disease progression tracking [1,2,10,15]. This step is both crucial and challenging as the deformations involved are large and may contain complex patterns such as sliding motion between organs. While traditional registration methods tend to fail on this task, learning approaches such as [5,6,12] obtained promising results at the Learn2Reg 2020 challenge, task 3 [7]. Yet, traditional and learning approaches both aims at

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-11203-4_9.

A. Hering et al. (Eds.): WBIR 2022, LNCS 13386, pp. 75–84, 2022.
https://doi.org/10.1007/978-3-031-11203-4_9

registering the whole image content instead of focusing on the relevant structures of interests. This introduces undesired noise and complexity to the registration process. To tackle this issue, we propose to exploit the recent availability of high quality automatic segmentation pipelines such as [3,17] and register the segmented structures. Specifically, structures are registered using their surface point cloud representation, allowing for exploiting meaningful geometric information of the different organs and finely modeling their dynamic properties. We also stress that surface point clouds are easy to derive from segmentation masks and are a lightweight representation of the structures of interest.

The Coherent Point Drift [13] (CPD) algorithm is one of the most popular method for deformable point cloud registration considered as state of the art [11]. A recent work [9] extended this framework using a Bayesian formulation and obtained more robust performances. CDP and BCPD both assume that points move coherently as a group to preserve the structure coherence. This is mainly because these frameworks are designed to register point clouds representing a single object. Consequently, [9,13] are not adapted for registering multi-organ points clouds. In particular, the coherency assumption doesn't stand for organs registration as each organ-specific point cloud may move independently to its neighbour, especially if we aim at registering inter-patient images.

In this work, we introduce a Multi-Organ Bayesian Coherent Point Drift algorithm (MO-BCPD) that models independent coherent structures. The contribution of this work is four-fold. Firstly, we extend the Bayesian formulation of CPD to model more complex structures interactions such as organ motion independence. Secondly, given that points clouds are obtained using automated segmentations, the proposed framework models partial segmentation errors allowing MO-BCPD to recover them. Thirdly, we model individual organ elasticity as part of the formulation. Fourthly, extensive experiments on 104 patients (10,712 pairs of patients) from the LiTS public dataset [4] demonstrate the effectiveness of our approach compared to BCPD. In particular, our method achieves an average target registration error on anatomical landmarks of 13 mm compared to 22 mm for the standard BCPD.

2 Method

In this section, we present our Multi-Organ Bayesian Coherent Point Drift algorithm. Let $\mathbf{y} = [\mathbf{y}_m]_{m \in \{1 \dots M\}} \in \mathbb{R}^{M,3}$ be the source point cloud and $\mathbf{x} = [\mathbf{x}_n]_{n \in \{1 \dots N\}} \in \mathbb{R}^{N,3}$ be the target point cloud where N and M are respectively the number of source and target points. We aim at finding the transformation $\mathcal{T}$ that realistically aligns these point clouds. In particular here, unlike in [9,13], the considered point clouds both represent a set of organ surfaces. Hence, each point is associated with an organ. Let $\mathbf{l}^y = [l^y_m]_{m \in \{1 \dots M\}} \in \{1 \dots L\}^M$ be the organ labels of the source point cloud and $\mathbf{l}^x = [l^x_n]_{n \in \{1 \dots N\}} \in \{1 \dots L\}^N$ be the organ labels of the target point cloud. L is the number of organs.

Transformation Model. Similarly to the BCPD, the Multi-Organ Bayesian Coherent Point Drift (MO-BCPD), decomposes the motion in two components: a sim-

Algorithm 1: Multi-Organ BCPD $(\mathbf{y}, \mathbf{x}, \omega, \Lambda, B, S, U, \kappa, \gamma, \epsilon)$

```
$\mathbf{v} \leftarrow 0_{M,3}$, $\Sigma \leftarrow Id_M$, $s \leftarrow 1$, $R \leftarrow Id_3$, $t \leftarrow 0_3$, $< \alpha_m > \leftarrow \frac{1}{M}$,
$\sigma^2 \leftarrow \frac{\gamma}{D \sum_{m,n} u_{l^y_m, l^x_n}} \sum_{m,n} u_{l^y_m, l^x_n} \|x_n - y_m\|^2$, $\theta \leftarrow (\mathbf{v}, \alpha, \mathbf{c}, \mathbf{e}, \rho, \sigma^2)$, $P \leftarrow \frac{1}{M} \mathbf{1}_{M,N}$
$\nu' \leftarrow 1_N$, $q_1(.,.) \leftarrow D^{\kappa \mathbf{1}_M} \phi^{0,\Sigma}$,
$q_2(c,e) \leftarrow \prod_{n=1}^{N} (1 - \nu'_n)^{1-c_n} \left( \nu'_n \prod_{m=1}^{M} \left( \frac{p_{mn}}{\nu'_n} \right)^{\delta_n(e_m)} \right)^{c_n}$, $q_3(.,.) \leftarrow \delta_\rho \delta_{\sigma^2}$
while $L(q_1 q_2 q_3)$ increases more than $\epsilon$ do
    Update $P$ and related terms:
    $\forall m, n \; \phi_{m,n} \leftarrow u_{l^y_m, l^x_n} \phi^{y'_m, \sigma^2 Id_3}(x_n) \exp - \frac{3 s^2 \Sigma_{m,m}}{2\sigma^2}$,
    $\forall m, n \; p_{m,n} \leftarrow \frac{(1-\omega) < \alpha_m > \phi_{m,n}}{\omega p_{out}(x_n) + (1-\omega) \sum_{m'} < \alpha_{m'} > \phi_{m',n}}$, $\nu \leftarrow P.1_N$, $\nu' \leftarrow P^T.1_M$,
    $\hat{N} \leftarrow \nu^T.1_M$, $\hat{\mathbf{x}} \leftarrow \Delta(\nu)^{-1}.P.\mathbf{x}$,
    Update displacement field and related terms:
    $\Sigma \leftarrow \left( G^{-1} + \frac{s^2}{\sigma^2} \Delta(\nu) \right)$, $\forall d \in \{1,2,3\} \; \mathbf{v}^d \leftarrow \frac{s^2}{\sigma^2} \Sigma \Delta(\nu) (\rho^{-1}(\hat{\mathbf{x}}^d) - \mathbf{y}^d)$,
    $\mathbf{u} \leftarrow \mathbf{y} + \mathbf{v}$, $< \alpha_m > \leftarrow exp\{\psi(\kappa + \nu_m) - \psi(\kappa M + \hat{N})\}$
    Update $\rho$ and related terms: $\bar{x} \leftarrow \frac{1}{\hat{N}} \sum_{m=1}^{M} \nu_m \hat{x}_m$, $\bar{\sigma}^2 \leftarrow \frac{1}{\hat{N}} \sum_{m=1}^{M} \nu_m \sigma^2_m$,
    $\bar{u} \leftarrow \frac{1}{\hat{N}} \sum_{m=1}^{M} \nu_m u_m$, $S_{xu} \leftarrow \frac{1}{\hat{N}} \sum_{m=1}^{M} (\hat{x}_m - \bar{x})(u_m - \bar{u})^T$,
    $S_{uu} \leftarrow \frac{1}{\hat{N}} \sum_{m=1}^{M} (u_m - \bar{u})(u_m - \bar{u})^T + \bar{\sigma}^2 Id_3$, $\Phi S'_{xu} \Psi^T \leftarrow svd(S_{xu})$,
    $R \leftarrow \Phi d(1, \ldots, 1, |\Phi\Psi|) \Psi^T$, $s \leftarrow \frac{Tr(R S_{xu})}{Tr(S_{uu})}$, $t \leftarrow \bar{x} - sR\bar{u}$, $\mathbf{y}' \leftarrow \rho(\mathbf{y} + \mathbf{v})$
    $\sigma^2 \leftarrow \frac{1}{3\hat{N}} \sum_{d=1}^{3} \left( (\mathbf{x}^d)^T \Delta(\nu') \mathbf{x}^d - 2\mathbf{x}^d P^T \mathbf{y}'^d + (\mathbf{y}'^d)^T \Delta(\nu) \mathbf{y}'^d \right) + s^2 \bar{\sigma}^2$
    Update q: $q_1(.,.) \leftarrow D^{\kappa \mathbf{1}_M} \phi^{\mathbf{v},\Sigma}$,
    $q_2(c,e) \leftarrow \prod_{j=1}^{N} (1 - \nu'_j)^{1-c_j} \left( \nu'_j \prod_{i=1}^{M} \left( \frac{p_{ij}}{\nu'_j} \right)^{\delta_i(e_j)} \right)^{c_j}$, $q_3(.,.) \leftarrow \delta_\rho \delta_{\sigma^2}$
end
```

ilarity transform $\rho : \mathbf{p} \longrightarrow s\mathbf{R}\mathbf{p} + \mathbf{t}$ and a dense displacement field $\mathbf{v}$. Hence the deformed source point could is $[\mathcal{T}(\mathbf{y}_m)]_{m \in \{1 \ldots M\}} = [\rho(\mathbf{y}_m + \mathbf{v}_m)]_{m \in \{1 \ldots M\}}$. While this parametrization is redundant, [9] has shown that this makes the algorithm more robust to target rotations. Moreover, it is equivalent to performing a rigid alignment followed by a non-rigid refinement which corresponds to the common practice in medical image registration.

Generative Model. As in [9], MO-BCPD assumes that all points from the target point cloud $[\mathbf{x}_n]_{n \in \{1 \ldots N\}}$ are sampled independently from a generative model. A point x_n from the target point cloud is either an outlier or an inlier which is indicated by a hidden binary variable c_n. We note the probability for a point to be an outlier ω (i.e. $\mathcal{P}(c_n = 0) = \omega$). If x_n is an outlier, it is sampled from an outlier distribution of density p_{out} (typically, a uniform distribution over a volume containing the target point cloud). If x_n is an inlier ($c_n = 1$), x_n is associated with a point $\mathcal{T}(y_m)$ in the deformed source point cloud. Let e_n be a multinomial variable indicating the index of the point of the deformed source point cloud with which x_n is associated (i.e. $e_n = m$ in our example). Let α_m be the probability of selecting the point $\mathcal{T}(y_m)$ to generate a point of the target point cloud (i.e. $\forall n \ \mathcal{P}(e_n = m|c_n = 1) = \alpha_m$). x_n is then sampled from a Gaussian distribution with covariance-matrix $\sigma^2 Id_3$ (Id_3 is the identity matrix of $\mathbb{R}^3$) centered on $\mathcal{T}(\mathbf{y}_m)$. Finally, the organ label l_n^x is sampled according to the label transition distribution $\mathcal{P}(l_n^x|l_m^y) = u_{l_n^x, l_m^y}$. The addition of the label transition term is our contribution to the original generative model [9]. This term encourages to map corresponding organs between the different anatomies while allowing to recover from partial segmentation errors from the automatic segmentation tool.

We can now write the following conditional probability density:

$$p^e(x_n, l_n^x, c_n, e_n|\mathbf{y}, \mathbf{l}^y, \mathbf{v}, \alpha, \rho, \sigma^2)$$
$$= (\omega p_{out}(x_n))^{1-c_n} \left((1-\omega) \prod_{m=1}^{M} \left(\alpha_m u_{l_m^y, l_n^x} \phi^{\mathbf{y}'_m, \sigma^2 Id_3}(\mathbf{x}_n) \right)^{\delta_{e_n = m}} \right)^{c_n} \quad (1)$$

where $\phi^{\mu,\Sigma}$ is the density of a multivariate Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ and δ is the Kronecker symbol.

Prior Distributions. MO-BCPD also relies on prior distributions in order to regularize the registration process and obtain realistic solutions. As in [9], MO-BCPD defines two prior distributions: $p^v(\mathbf{v}|\mathbf{y}, \mathbf{l}_y)$ that regularizes the dense displacement field and $p^\alpha(\alpha)$ that regularizes the parameters α of the source point cloud selection multinomial distribution mentioned in the generative model. The prior on α follows a Dirichlet distribution of parameter $\kappa \mathbf{1}_M$. In practice, κ is set to a very high value which forces $\alpha_m \approx 1/M$ for all m. To decouple motion characteristics within and between organs, we propose a novel formulation of the displacement field prior p^v. Specifically, we introduce 3 parameters: a symmetric matrix $S = [s_{l,l'}]_{l,l' \in \{0 \dots L\}}$ and two vectors $\Lambda = [\Lambda_l]_{l \in \{0 \dots L\}}$ and $B = [B_l]_{l \in \{0 \dots L\}}$. The matrix S parametrizes the motion coherence inter-organs. The vectors Λ and B respectively characterizes the variance of the deformation magnitude and motion coherence bandwidth within each organ. We define the displacement field prior for the MO-BCPD as:

$$p^v(\mathbf{v}|\mathbf{y}, \mathbf{l}^y) = \phi^{\mathbf{0},G}(\mathbf{v}^1)\phi^{\mathbf{0},G}(\mathbf{v}^2)\phi^{\mathbf{0},G}(\mathbf{v}^3) \quad (2)$$

$$G = \left[\Lambda_{l_i^y} \Lambda_{l_j^y} S_{l_i^y, l_j^y} \exp - \frac{\|y_i - y_j\|^2}{2 B_{l_i^y} B_{l_j^y}} \right]_{i,j \leq M} \quad (3)$$

Note that G must be definite-positive, leading to strictly positive values for variance of the displacement magnitude Λ_l and mild constraints on S.

Learning. Combining equations (1) and (2), the joint probability distribution of the variables $\mathbf{y}, \mathbf{l}^y, \mathbf{x}, \mathbf{l}^x, \theta$, where $\theta = (\mathbf{v}, \alpha, \mathbf{c}, \mathbf{e}, \rho, \sigma^2)$ is defined as:

$$p(\mathbf{x}, \mathbf{l}^x, \mathbf{y}, \mathbf{l}^y, \theta) \propto p^v(\mathbf{v}|\mathbf{y}, \mathbf{l}_y) p^\alpha(\alpha) \prod_{n=1}^{N} p^e(x_n, l_n^x, c_n, e_n|\mathbf{y}, \mathbf{l}_y, \mathbf{v}, \alpha, \rho, \sigma^2) \quad (4)$$

As in [9], we use variational inference to approximate the posterior distribution $p(\theta|\mathbf{x}, \mathbf{y})$ with a factorized distribution $q(\theta) = q_1(\mathbf{v}, \alpha)q_2(\mathbf{c}, \mathbf{e})q_3(\rho, \sigma^2)$ so that $q = \underset{q_1,q_2,q_3}{\arg\min}\, KL(q|p(.|\mathbf{x}, \mathbf{y}))$ where KL is the Kullback-Leibler divergence. Similarly to [9], we derive the MO-BCPD algorithm presented in algorithm 1. The steps detailed in Algorithm 1 perform coordinate ascent on the evidence lower bound $L(\theta) = \int_\theta q(\theta) \ln \frac{p(\mathbf{x},\mathbf{y},\theta)}{q(\theta)} d\theta$. In algorithm 1, γ is a hyper-parameter used to scale the initial estimation of σ^2 and ϵ is used for stopping criteria. We note $\Delta(\nu)$ the diagonal matrix with diagonal entries equal to ν.

Hyper-parameter Setting. The model has a large number of hyper-parameters which can impact the performance of the algorithm. Regarding κ, the parameter of the prior distribution p^α, and γ, the scaling applied to the initial estimation of σ^2, we followed the guidelines in [9]. ω is set based on an estimate of the proportion of outliers on a representative testing set. Regarding B and Λ, respectively the vector of organ-specific motion coherence bandwidth and expected deformation magnitude, they characterise organs elastic properties. Concretely, a larger motion coherence bandwidth B_l increases the range of displacement correlation for organ l (points that are further away are encouraged to move in the same direction). A larger expected deformation magnitude Λ_l increases the probability of larger displacements for organ l. These are physical quantities expressed in mm that could be set based on organs physical properties. The inter-organ motion coherence matrix S should be a symmetric matrix containing values between 0 and 1. $S_{l,l} = 1$ for all organs $l \in \{1 \dots L\}$ and $S_{l,l'}$ is closer to 0 if organs l and l' can move independently.

$u_{l,l'}$ is the probability that a point with label l' generates a point with label l. As points labels are in practice obtained from an automatic segmentation tool, we note $[g^y_m]_{m\in\{1...M\}}$ and $[g^x_n]_{n\in\{1...n\}}$ respectively the unknown true organ labels of the source and target point clouds (as opposed to the estimated ones $[l^y_m]_{m\in\{1...M\}}$ and $[l^y_m]_{m\in\{1...M\}}$). We assume that points from the deformed source point cloud generate points with the same true labels (i.e. $\mathcal{P}(g^x_n = g^y_m|e_n = m) = 1$). Hence, the probability for y_m, with estimated organ label l^y_m to generate a point with label l^x_n is given by: $u_{l^x_n,l^y_m} = \sum_k p(g^y_m = k|l^y_m)p(l^x_n|g^x_n = k)$ where $p(g^y_m = k|l^y_m)$ is the probability that a point labelled l^y_m by the automatic segmentation tool has true label k and $p(l^x_n|g^x_n = k)$ is the probability that the automatic segmentation tool predicts the organ label l^x_n for a point with true label k. These probabilities need to be estimated on a representative testing set. We note that if the segmentation is error-free, the formula above gives $U = Id_L$. Indeed, in that case the points organ labels correspond exactly to the organ true labels so a point belonging to a certain organ can only generate a point from the same organ in the target anatomy. Properly setting the organ label transition probability matrix U is crucial to recover from potential partial segmentation errors. Figure 1 illustrates with a toy example a situation where the algorithm converges to an undesired state if the segmentation error is not modeled properly.

Interpolation. Once the deformation on the organ point clouds is known, one might want to interpolate the deformation back to image space in order to

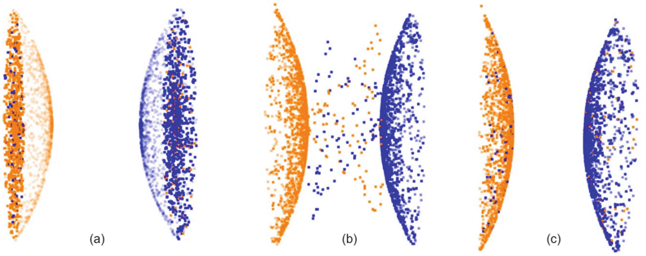


Fig. 1. Toy example registering a pair of organs (a blue and an orange organ) with $\sim$10% segmentation error (corrupted input labels). Both organs (orange and blue) of the target point cloud are shown in (a) in transparent while the source point cloud is shown in opaque. The blue (orange) dots on the left (right) of the figure corresponds to simulated segmentation errors. (b) shows the registered point cloud without modeling the inter-organ segmentation error, (c) shows the registered point cloud with segmentation error modelization

resample the whole volume. As in [8] we propose to use Gaussian process regression to interpolate the deformation obtained by the MO-BCPD algorithm. This interpolation process can also be used to register sub-sampled point clouds to decrease computation time as in [8].

Given a set of points $\tilde{\mathbf{y}} = [\tilde{y}_i]_{i \in \{1 \ldots \tilde{M}\}}$ with labels $\mathbf{l}^{\tilde{y}} = [l_i^{\tilde{y}}]_{i \in \{1 \ldots \tilde{M}\}}$. We compute the displacement for the set of points $\tilde{y}$ as:

$$\mathbf{v}^{\tilde{\mathbf{y}}} = G^{int}(\tilde{\mathbf{y}}, \mathbf{l}^{\tilde{y}}, \mathbf{y}, \mathbf{l}^{y}, B, \Lambda, S).G^{-1}.v \tag{5}$$

$$G^{int}(\tilde{\mathbf{y}}, \mathbf{l}^{\tilde{y}}, \mathbf{y}, \mathbf{l}^{y}, B, \Lambda, S)_{i,j} = \Lambda_{l_i^{\tilde{y}}} \Lambda_{l_i^{y}} S_{l_i^{\tilde{y}}, l_i^{y}} \exp - \frac{\|\tilde{y}_i - y_j\|^2}{2\beta_{l_i^{\tilde{y}}} \beta_{l_j^{y}}} \tag{6}$$

Acceleration. The speed ups strategies mentioned in [9] are fully transferable to the MO-BCPD pipeline. In our experiments though, the main improvement, by far, came from performing a low rank decomposition of G at the initialization of the algorithm. Indeed, this yielded consistent reliable ×10 speed-ups with negligible error when using ≥ 20 eigen values. The Nystrom methods to approximate P sometimes implied large error due to the stochasticity of the method while yielding up to ×2 speed-ups which is why we did not use it. This allows MO-BCPD to be run in a few seconds with $M, N \approx 5000$.

3 Experiments

We evaluate the MO-BCPD algorithm by performing inter-patient registration from the LITS challenge training dataset [4] which contains 131 chest-CT patient images. 27 patients were removed due to different field of view, issues with the segmentation or landmark detection. In total, 10,712 registrations were performed on all the pairs of remaining patients. The segmentation was automatically performed using an in-house tool derived from [17] which also provides a

Table 1. Target registration error on landmarks. Results in mm (std).

	Sim	BCPD	GMC-MO-BCPD	OMC-MO-BCPD
Bladder	257 (26)	29 (15)	30 (15)	**26** (15)
Left kidney bottom	128 (18)	23 (11)	22 (10)	**8** (4)
Left kidney center	107 (13)	18 (10)	15 (8)	**6** (3)
Left kidney top	101 (15)	23 (12)	21 (10)	**9** (4)
Liver bottom	114 (15)	29 (14)	28 (14)	**24** (13)
Liver center	65 (10)	12 (7)	12 (7)	**11** (7)
Liver top	123 (15)	**24** (12)	25 (12)	26 (14)
Right kidney bottom	98 (15)	26 (13)	24 (11)	**10** (6)
Right kidney center	65 (11)	21 (12)	17 (9)	**5** (3)
Right kidney top	64 (15)	25 (13)	21 (11)	**9** (4)
Round ligament of liver	95 (20)	27 (14)	27 (13)	**25** (13)

set of anatomical landmarks for each image which were used for evaluation. We considered five organs of interest: the liver, the spleen, the left and right kidneys and the bladder. We compared 4 different algorithms: registration of the point clouds with a similarity transform (Sim), BCPD, GMC-MO-BCPD which is MO-BCPD with global motion coherence ($S = 1_{L,L}$) and OMC-MO-BCPD which is MO-BCPD with intra-organ motion coherence only ($S = Id_L$). As the segmentation tool performed very well on the considered organs, we set $U = Id_L$ and $\omega = 0$ for both MO-BCPD versions ($\omega = 0$ for BCPD as well). We used for all organs the same values for Λ_l and B_l respectively 10 mm and 30 mm as a trade off between shape matching and preservation of individual organs appearance ($\beta = 30$ and $\lambda = 0.1$ for BCPD which is the equivalent configuration). We also set, $\gamma = 1$ and $\epsilon = 0.1$. We compared those algorithms by computing the registration error on the anatomical landmarks belonging to those organs. We chose this generic, relatively simple setting (same rigidity values for all organs, no outlier modeling, only two extreme configurations for S) in order to perform large scale inter-patient registration experiments but we would like to stress that further fine tuning of these parameters for a specific application or even for a specific patient would further improve the modeling and hence the registration outcome. Results are presented in Table 1. We observe that while GMC-MO-BCPD induces some marginal improvements with respect to BCPD, OMC-MO-BCPD allows a much more precise registration. As illustrated in Fig. 2, the main improvement from BCPD to GMC-MO-BCPD is that different organs no longer overlap. Indeed, as highlighted by the green and blue ellipses, the spleen and the right kidney from the deformed source patient overlap the liver of the target patient when using BCPD. OMC-MO-BCPD properly aligns the different organs while preserving their shape (see orange and purple ellipses for instance).

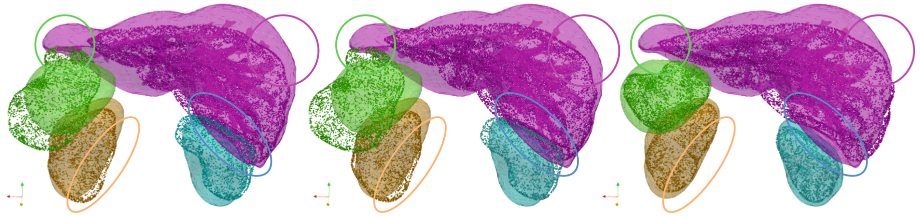


Fig. 2. Qualitative comparison of registration output for the liver, left/right kidneys and spleen. From left to right, BCPD, GMC-MO-BCPD, OMC-MO-BCPD. The target organs are shown in transparency while the deformed point cloud are shown in opaque.

4 Conclusion

We introduced MO-BCPD, an extension of the BCPD algorithm specifically adapted to abdominal organ registration. We identified three limitations of the original work [9] on this task and proposed solutions to model: the segmentation error between neighboring organs of interest, the heterogeneous elastic properties of the abdominal organs and the complex interaction between various organs in terms of motion coherence. We demonstrated significant improvements over BCPD on a large validation set (N=10,712).

Moreover, we would like to highlight that segmentation error could also be taken into account by tuning the outlier probability distribution p_{out} and the probability of being an outlier ω. When the point is estimated as a potential outlier by the algorithm, its contribution to the estimation of the transformation $\mathcal{T}$ is lowered. Hence, the segmentation error modelled by p_{out} and ω corresponds to over/under segmentation, i.e. when there is a confusion between an organ and another class we don't make use of in MO-BCPD (e.g. background). Hence, MO-BCPD introduces a finer way of handling segmentation error by distinguishing two types of errors: mis-labeling between classes of interest which is modelled by U and over/under-segmentation of classes of interest modeled by ω and p_{out}.

In this manuscript, we focused on highlighting the improvements yielded by the MO-BCPD formulation specifically designed for multi-organ point cloud registration. That being said, some clinical applications would require the deformation on the whole original image volume. Hence, we present in supplementary material preliminary results on a realistic clinical use case. In Figs. 3 and 4 we see that MO-BCPD coupled with the proposed interpolation framework obtain better alignment on the structures of interest than traditional intensity-based baselines. It is also interesting to note that MO-BCPD also better aligns structures that are particularly challenging to align such as the hepatic vein while not using these structures in the MO-BCPD.

Future work will investigate how MO-BCPD could be used as a fast, accurate initialization for image-based registration algorithms. From a modeling standpoint, we would also like to further work on segmentation error modeling (in particular over/under segmentation) with a more complex organ specific outlier distribution.

References

1. Carrillo, A., Duerk, J., Lewin, J., Wilson, D.: Semiautomatic 3-D image registration as applied to interventional MRI liver cancer treatment. IEEE Trans. Med. Imaging **19**(3), 175–185 (2000). https://doi.org/10.1109/42.845176
2. Cash, D.M., et al.: Concepts and preliminary data toward the realization of image-guided liver surgery. J. Gastrointest. Surg. **11**(7), 844–59 (2007). https://www.proquest.com/scholarly-journals/concepts-preliminary-data-toward-realization/docview/1112236808/se-2?accountid=11862. copyright - The Society for Surgery of the Alimentary Tract 2007; Dernière mise á jour - 2014–03-30
3. Chen, X., et al.: A deep learning-based auto-segmentation system for organs-at-risk on whole-body computed tomography images for radiation therapy. Radiother. Oncol. **160**, 175–184 (2021). https://doi.org/10.1016/j.radonc.2021.04.019
4. Christ, P.F., et al.: The liver tumor segmentation benchmark (LiTS). CoRR abs/1901.04056 (2019). http://arxiv.org/abs/1901.04056
5. Estienne, T., et al.: Deep learning based registration using spatial gradients and noisy segmentation labels. CoRR abs/2010.10897 (2020). https://arxiv.org/abs/2010.10897
6. 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
7. Hering, A., et al.: Learn2Reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning. arXiv preprint arXiv:2112.04489 (December 2021)
8. Hirose, O.: Acceleration of non-rigid point set registration with downsampling and Gaussian process regression. IEEE Trans. Pattern Anal. Mach. Intell. **43**(8), 2858–2865 (2021). https://doi.org/10.1109/TPAMI.2020.3043769
9. Hirose, O.: A Bayesian formulation of coherent point drift. IEEE Trans. Pattern Anal. Mach. Intell. **43**(7), 2269–2286 (2021). https://doi.org/10.1109/TPAMI.2020.2971687
10. Lange, T., et al.: Registration of portal and hepatic venous phase of MR/CT data for computer-assisted liver surgery planning. In: International Congress Series, vol. 1281, pp. 768–772 (2005). https://doi.org/10.1016/j.ics.2005.03.332
11. Maiseli, B., Gu, Y., Gao, H.: Recent developments and trends in point set registration methods. J. Vis. Commun. Image Represent. **46**, 95–106 (2017). https://doi.org/10.1016/j.jvcir.2017.03.012, https://www.sciencedirect.com/science/article/pii/S1047320317300743
12. Mok, T.C.W., Chung, A.C.S.: Large deformation diffeomorphic image registration with Laplacian pyramid networks. ArXiv abs/2006.16148 (2020)
13. Myronenko, A., Song, X.: Point set registration: coherent point drift. IEEE Trans. Pattern Anal. Mach. Intell. **32**(12), 2262–2275 (2010). https://doi.org/10.1109/TPAMI.2010.46
14. Papież, B.W., Franklin, J.M., Heinrich, M.P., Gleeson, F.V., Brady, M., Schnabel, J.A.: Gifted demons: deformable image registration with local structure-preserving regularization using supervoxels for liver applications. J. Med. Imaging **5**, 024001 (2018). https://doi.org/10.1117/1.JMI.5.2.024001
15. Robu, M.R., et al.: Global rigid registration of CT to video in laparoscopic liver surgery. Int. J. Comput. Assist. Radiol. Surg. **13**, 947–956 (2018)

16. Thirion, J.P.: Image matching as a diffusion process: an analogy with Maxwell's demons. Med. Image Anal. **2**(3), 243–260 (1998). https://doi.org/10.1016/S1361-8415(98)80022-4, https://www.sciencedirect.com/science/article/pii/S1361841598800224
17. Yang, D., et al.: Automatic liver segmentation using an adversarial image-to-image network. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 507–515. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66179-7_58

Voxelmorph++

Going Beyond the Cranial Vault with Keypoint Supervision and Multi-channel Instance Optimisation

Mattias P. Heinrich(✉) and Lasse Hansen

Institute of Medical Informatics, University of Lübeck, Lübeck, Germany
{heinrich,hansen}@imi.uni-luebeck.de

Abstract. The majority of current research in deep learning based image registration addresses inter-patient brain registration with moderate deformation magnitudes. The recent Learn2Reg medical registration benchmark has demonstrated that single-scale U-Net architectures, such as VoxelMorph that directly employ a spatial transformer loss, often do not generalise well beyond the cranial vault and fall short of state-of-the-art performance for abdominal or intra-patient lung registration. Here, we propose two straightforward steps that greatly reduce this gap in accuracy. First, we employ keypoint self-supervision with a novel network head that predicts a discretised heatmap and robustly reduces large deformations for better robustness. Second, we replace multiple learned fine-tuning steps by a single instance optimisation with hand-crafted features and the Adam optimiser. Different to other related work, including FlowNet or PDD-Net, our approach does not require a fully discretised architecture with correlation layer. Our ablation study demonstrates the importance of keypoints in both self-supervised and unsupervised (using only a MIND metric) settings. On a multi-centric inspiration-exhale lung CT dataset, including very challenging COPD scans, our method outperforms VoxelMorph by improving nonlinear alignment by 77% compared to 19% - reaching target registration errors of 2 mm that outperform all but one learning methods published to date. Extending the method to semantic features sets new stat-of-the-art performance on inter-subject abdominal CT registration.

Keywords: Registration · Heatmaps · Deep learning

1 Introduction

Medical image registration aims at finding anatomical and semantic correspondences between multiple scans of the same patient (intra-subject) or across a population (inter-subject). The difficulty of this task with plentiful clinical applications lies in discriminating between changes in intensities due to image appearance changes (acquisition protocol, density difference, contrast, etc.) and nonlinear deformations. Advanced similarity metrics may help in finding a good

A. Hering et al. (Eds.): WBIR 2022, LNCS 13386, pp. 85–95, 2022.
https://doi.org/10.1007/978-3-031-11203-4_10

contrast-invariant description of local neighbourhoods, e.g. normalised gradient fields [7] or MIND [14]. Due to the ill-posedness of the problem some form of regularisation is often employed to resolve the disambiguity between several potential local minima in the cost function. Powerful optimisation frameworks that may comprise iterative gradient descent, discrete graphical models or both (see [28] for an overview) aim at solving for a global optimum that best aligns the overall scans (within the respective regions of interest). Many deep learning (DL) registration frameworks (e.g. DLIR [30] and VoxelMorph [1]) rely on a spatial transformer loss that may be susceptible to ambiguous optimisation landscapes - hence multiple resolution or scales levels need to be considered. The focus of this work is to reflect such local minima more robustly in the loss function of DL-registration. We propose to predict probabilistic displacement likelihoods as heatmaps, which can better capture multiple scales of deformation within a single feed-forward network.

Related Work: Addressing large deformations with learning based registration is generally approached by either multi-scale, label-supervised networks [15,21,22] or by employing explicitly discretised displacements [9,11]. Many variants of U-Net like architectures have been proposed that include among others, dual-stream [18], cascades [32] and embeddings [5]. Different to those works, we do neither explicitly model a discretised displacement space, multiple scales or warps, nor modify the straightforward feed-forward U-Net of DLIR or VoxelMorph. Other works aimed at learning lung deformations through simulated transformations [4,26].

Learning based point-cloud registration is another research field of interest (FlowNet3d [20]) that has however so far been restricted to lung registration in the medical domain [8]. For a comparison of classical approaches using thin-plate splines or expectation-maximisation e.g. coherent point drift to learning-based ones the reader is referred to comparison experiments in [8]. Stacked hourglass networks that predict discretised heatmaps of well-defined anatomical landmarks are commonly used in human pose estimation [24]. They are, however, restricted to datasets and registration applications where not only pairwise one-to-one correspondences can be obtained as training objective but generic landmarks have to be found across all potential subjects. Due to anatomical variations this restriction often prevents their use in medical registration. DRAMMS [25] also aims at matching keypoints across scans and learns discriminative features based on mutual saliency maps but does not offer the benefits of fast feed-forward prediction of displacements.

Combining an initial robust deformation prediction with instance optimisation or further learning fine-tuning steps (Learn-to-optimise, cf. [29]) has become a new trend in learning-based registration, e.g. [16,27]. This paradigm, which expects coarse but large displacements from a feed-forward prediction, can shift the focus away from sub-pixel accuracy and towards avoiding failure cases in the global transformation due to local minima. It is based on the observation that local iterative optimisers are very precise when correctly initialised and have become extremely fast due to GPU acceleration.

Contributions: We demonstrate that a single-scale U-Net without any bells and whistles in conjunction with a fast MIND-based instance optimisation can achieve or outperform state-of-the-art in large-deformation registration. This is achieved by focussing on coarse scale global transformation by introducing a novel heatmap prediction network head. In our first scenario we employ weak self-supervision, through automatic keypoint correspondences [12]. Here, the heatmap enables a discretised integral regression of displacements to directly and explicitly match the keypoint supervision. Second, we incorporate the heatmap prediction into a non-local unsupervised metric loss. This enables a direct comparison within the same network architecture to the commonly used spatial transformer (warping) loss in unsupervised DL registration and highlights the importance of providing better guidance to avoid local minima. Our extensive ablation experiments with and without instance optimisation on a moderately large and challenging inspiration-exhale lung dataset demonstrate state-of-the-art performance.

Our code is publicly available at: https://www.github.com/mattiaspaul/VoxelMorphPlusPlus

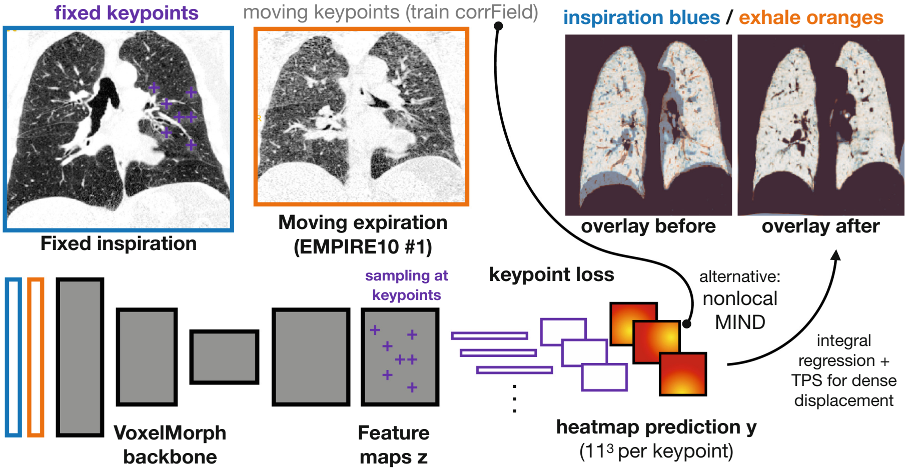

fixed keypoints
moving keypoints (train corrField)
inspiration blues / exhale oranges
Fixed inspiration
Moving expiration (EMPIRE10 #1)
overlay before
overlay after
sampling at keypoints
keypoint loss
alternative: nonlocal MIND
integral regression + TPS for dense displacement
VoxelMorph backbone
Feature maps z
heatmap prediction y
(11³ per keypoint)

Fig. 1. Overview of method and qualitative result for held-out case #1 from [23]. The key new element of our approach is the heatmap prediction head that is appended to a standard VoxelMorph. It helps overcome local minima through a probabilistic loss with either automatic keypoint correspondences or non-locally weighted MIND features.

2 Method

Keypoint correspondences are an excellent starting point to explore the benefits of incorporating a heatmap prediction within DL-registration. Our method can be either trained when those automatically computed displacements at around $|K| \approx 2000$ locations per scan are available for training data, or we can use a non-local unsupervised metric loss (see details below). In both scenarios, we use Förstner keypoints in the fixed scan to focus on distinct locations within a region of interest. We will first describe the baseline single-scale U-Net backbone, followed by our novel heatmap prediction head and the non-local MIND loss (which is an extension of [13] to 3D).

Baseline Backbone: Given two input CT scans, fixed and moving image $F, M : \mathbb{R}^3 \rightarrow \mathbb{R}$ and a region of interest $\Omega \in \mathbb{R}^3$, we firstly define a feed-forward U-Net [6] $\Theta(F, M, \Omega, \theta) \rightarrow \mathbb{R}^C$ with trainable parameters θ that maps the concatenated input towards a shared C-dimensional feature representation $\mathbf{z}$ (we found $\mathrm{C} \approx 64$ is expressive enough to represent displacements). This representation may have a lower spatial resolution than F or M and is the basis for predicting a (sparse) displacement field φ that spatially aligns F and M within Ω. Θ comprises in our implementation a total of eleven 3D convolution blocks, each consisting of a $3 \times 3 \times 3$ convolution, instance normalisation (IN), ReLU, a $1 \times 1 \times 1$ convolution, and another IN+ReLU. Akin to VoxelMorph, we use $2 \times 2 \times 2$ max-pooling after each of the four blocks in the encoder and nearest neighbour upsampling to restore the resolution in the decoder, but use a half-resolution output. The network has up to $C = 64$ hidden feature channels and 901'888 trainable parameters.

Due to the fact that this backbone already contains several convolution blocks on the final resolution at the end of the decoder, it is directly capable of predicting a continuous displacement field φ by simply appending three more $1 \times 1 \times 1$ convolutions (and IN + ReLU) with a number of output channels equal to 3.

Discretised Heatmap Prediction Head: The aim of the heatmap prediction head is to map a C-dimensional feature vector (interpreted as a $1 \times 1 \times 1$ spatial tensor with $|K|$ being the batch dimension) into a discretised displacement tensor $y \in \mathcal{Q}$ with predefined size and spatial range R (see Fig. 1). Here we chose $R = 0.3$, in a coordinate system that ranges from -1 to $+1$, which captures even large lung motion between respiratory states. We define $\mathcal{Q}$ to be a discretised map of size $11 \times 11 \times 11$ to balance computational complexity and limit quantisation effects. This means we need to design another *nested* decoder that increases the spatial resolution from 1 to 11. Our heatmap network comprises a transpose 3D convolution with kernel size $7 \times 7 \times 7$, six further 3D convolution blocks (kernel size $7 \times 7 \times 7$ and IN+ReLU) once interleaved with a single trilinear upsampling to $11 \times 11 \times 11$. It has 462'417 trainable parameters and its number of output channels is equal to 1.

Next, we can define a probabilistic displacements tensor $\mathcal{P}$ with a dimensionality of 6 (3 spatial and 3 displacement dimensions) using a softmax operation along the combined displacement dimensions as:

$$\mathcal{P}(\mathbf{x}, \Delta\mathbf{x}) = \frac{\exp(y(\mathbf{x}, \Delta\mathbf{x}))}{\sum_{\Delta\mathbf{x}} \exp(y(\mathbf{x}, \Delta\mathbf{x}))}, \tag{1}$$

where $\mathbf{x}$ are global spatial 3D coordinates and $\Delta\mathbf{x}$ local 3D displacements. In order to define a continuous valued displacement field, we apply a weighted sum:

$$\varphi(\mathbf{x}) = \sum_{\Delta\mathbf{x}} \mathcal{P}(\mathbf{x}, \Delta\mathbf{x}) \cdot \mathcal{Q}(\Delta\mathbf{x}) \tag{2}$$

This output is used during training to compute a mean-squared error between predicted and pre-computed keypoint displacements. Since, the training correspondences are regularised using a graphical model, we require no further penalty.

Non-local MIND Loss: To avoid the previously described pitfalls of directly employing a spatial transformer (warping) loss, we can better employ the probabilistic heatmap prediction and compute the discretely warped MIND vectors of the moving scan implicitly by a weighted average of the underlying features within pre-defined capture region (where c describes one of the 12 MIND channels) as:

$$MIND_{warped}(c, \mathbf{x}) = \sum_{\Delta\mathbf{x}} \mathcal{P}(\mathbf{x}, \Delta\mathbf{x}) \cdot MIND(c, \mathbf{x} + \Delta\mathbf{x}) \tag{3}$$

.

Implementation Details: Note that the input to both the small regression network (baseline) and our proposed are feature vectors sampled at the keypoint locations, which already improves the baseline architecture slightly. We use trilinear interpolation in all cases where the input and output grids differ in size to obtain off-grid values. All predicted sparse displacements φ are extrapolated to a dense field using thin-plate-splines with $\lambda = 0.1$ that yields φ^*.

For the baseline regression setup (VoxelMorph) we employ a common MIND warping loss and a diffusion regularisation penalty that is computed based on the Laplacian of a kNN-graph ($k = 7$) between the fixed keypoints. The weighting of the regularisation was empirically set to $\alpha = 0.25$. We found that using spatially aggregated CT and MIND tensors the former using average pooling with kernel size 2, the latter two of those pooling steps, leads to stabler training in particular for the regression baseline.

Multi-channel Instance Optimisation: We directly follow the implementation described in [27][1]. It is initialised with φ^*, runs for 50 iterations, employs a combined B-spline and diffusion regularisation coupled with a MIND metric loss and a grid spacing of 2. This step is extremely fast, but relies on a robust initialisation as we will demonstrate in our experiments. The method can also be employed when semantic features, e.g. segmentation predictions from an nnUNet [19], are available in the form of one-hot tensors.

3 Experiments and Results

We perform extensive experiments on inspiration-exhale registration of lung CT scans - arguably one of the most challenging tasks in particular for learning-based

[1] https://github.com/multimodallearning/convexAdam.

registration [16]. A dataset of 30 scan pairs with large respiratory differences is collected from EMPIRE10 (8 scan pairs #1, #7, #8, #14, #18, #20, #21 and #28) [23], Learn2Reg Task 2 [17] and DIR-Lab COPD [2] (10 pairs). The exhale and inspiration scans are resampled to 1.75 × 1.25 × 1.75 mm and 1.75 × 1.00 × 1.25 mm respectively to account for average overall volume scaling and a fixed region with dimensions 192 × 192 × 208 voxels was cropped that centres the mass of automatic lung masks. Note that this pre-processing approximately halves the initial target registration error (TRE) of the COPD dataset. Lung masks are also used to define a region-of-interest for the loss evaluation and to mask input features for the instance optimisation. We split the data into five folds for cross-validation that reflect the multi-centric data origin (i.e. approx. two scans per centre are held out for validation each).

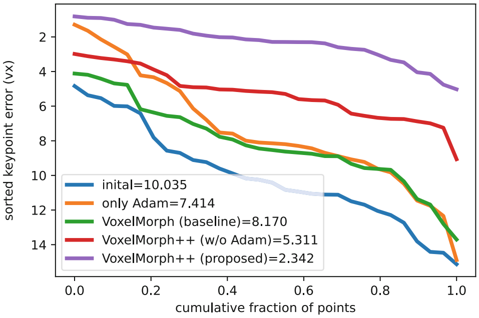


Fig. 2. Cumulative keypoint error of proposed model compared to a VoxelMorph baseline and using only Adam instance optimisation with MIND.

Table 1. Results of ablation study on lung CT: VoxelMorph++ improves error reduction of nonlinear alignment from 18% to 77%.

	UNet	Heatmap	Keypoints	Error w/o Adam	Error w/Adam
Initial/Adam				10.04 vx	7.41 vx
VoxelMorph	✓			**8.17 vx**	4.80 vx
VM + Heatmap	✓	✓		6.49 vx	3.18 vx
VM + Keypoints	✓		✓	6.30 vx	2.79 vx
VoxelMorph++	✓	✓	✓	5.31 vx	**2.34 vx**

Keypoint Self-supervision: To create correspondence as self-supervision for our proposed VoxelMorph++ method, we employ the **corrField** [12][2], which is designed for lung registration and based on a discretised displacement search and a Markov random field optimisation with multiple task specific improvements. It runs within a minute per scan pair and creates ≈2000 = $|K|$ highly accurate (≈1.68 mm) correspondences at Förstner keypoints.

As additional experiments we also apply the same technique to the popular (but less demanding) DIR-Lab 4DCT lung CT benchmark and we extend our method to the inter-subject alignment of 30 abdominal CTs [31] that was also part of the Learn2Reg 2020 challenge (Task 3) and provides 13 difficult anatomical organ labels for training and validation.

Ablation Study: We consider a five-fold cross-validation for all ablation experiments with an initial error of 10.04 vx (after translation and scaling) across 30 scan pairs computed based on keypoint correspondences. Employing only the

[2] http://www.mpheinrich.de/code/corrFieldWeb.zip.

Adam instance optimisation with MIND features results in an error of 8.17 vx, with default settings of grid spacing = 2 voxels, 50 iterations and $\lambda_{Adam} = 0.65$. Note that a dense displacement is estimated with a parametric B-spline model. We start from the slightly improved VoxelMorph **baseline** with MIND loss, diffusion regularisation and increased number of convolution operations described above. This yields a keypoint error of **8.17 vx** that represents an error reduction of 19% and can be further improved to 4.80 vx when adding instance optimisation. A weighting parameter $\lambda = 0.75$ for diffusion regularisation was empirically found with $k = 7$ for the sparse neighbourhood graph of keypoints. Replacing the traditional spatial transformer loss with our proposed heatmap prediction head that uses the nonlocal MIND loss much improves the performance to 6.49 vx and 3.18 vx (with and without Adam respectively). But the key improvement can be gained when including the self-supervised keypoint loss. Using our baseline VoxelMorph architecture that regresses continuous 3D vectors, we reach 6.30 and **2.79 vx**. See Table 1 and Fig. 2 for numerical and cumulative errors. Our heatmap-based network and the instance optimisation require around 0.43 and 0.41 s inference time, respectively. The complexity of transformations measured as standard deviation of log-Jacobian determinants is on average 0.0554. The number of negative values is zero (no folding) in 8 out of 10 COPD cases and negligible ($<10^{-4}$) in the others.

Comparison to State-of-the-Art on DIR-Lab: When evaluation the target registration error (TRE) in mm for the 10 pairs of DIR-Lab COPD [2] one of the most challenging benchmarks in medical registration with an initial misalignment of 23.36 mm (and 12.0 mm after pre-alignment), we reach 2.16 mm. This compares very favourable to VoxelMorph+ with 7.98 mm and LapIRN with 3.68 mm (see Table 2). Of all published DL-methods only GraphRegNet [9] is superior with 1.34 mm. The high visual quality of our registration is shown in Fig. 3.

Table 2. Target registration error in mm for 10 pairs each of DIR-Lab 4D lung-CT [3] and COPD [2] datasets in comparison to a selection of other published methods.

Method (citation)	Before	[26]	[15]	[4]	[21][a]	[1]	[9]	Ours
TRE (4DCT)	8.46 ± 6.6	2.52 ± 3.0	1.14 ± 0.8	3.68 ± 3.3	1.60	1.71 ± 2.9	1.39 ± 1.3	**1.33**
TRE (COPD)	23.36 ± 11.9	–	–	–	3.83	7.98 ± 3.8	1.34 ± 1.3	**2.16**

[a] own experiments including instance optimisation

Limitations and Further Potential: We have not yet considered more advanced network architectures as backbone, e.g. two-stream or multi-level, which are likely to yield further improvements. However, based on our experiments we expect that it could merely reduce the reliance on instance optimisation.

Inter-subject Abdominal CT Registration: We apply our proposed VoxelMorph++ model with nonlocal loss and no architectural modifications to

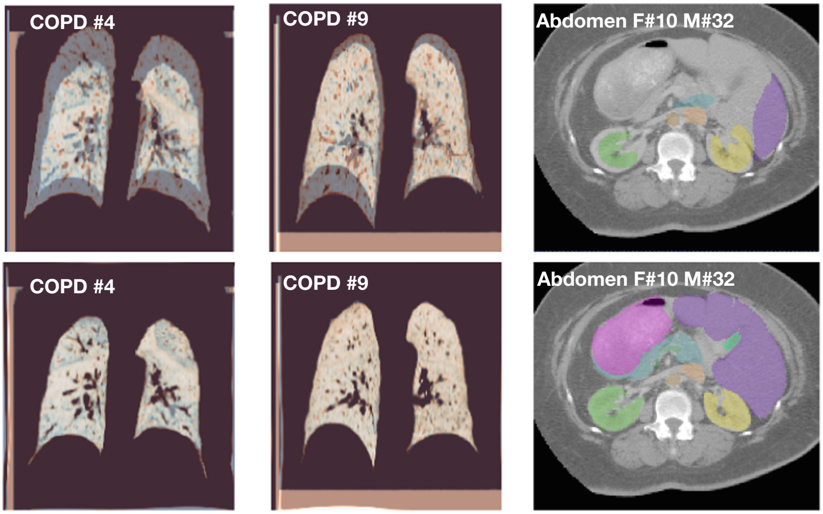


Fig. 3. Exemplary results of our proposed method before (top row) and after (bottom row) registration. Fixed exhale scans are shown in blue and inspiration in orange shades respectively (adding up to grayscale when aligned). For abdominal alignment transformed segmentation labels are shown, here: right kidney 🟨, left kidney 🟩, gallbladder 🟩, liver 🟪, stomach 🟪, aorta 🟫 and pancreas 🟦 are visible. (Color figure online)

another challenging task of inter-subject abdominal CT registration with initial Dice overlap for 13 organs of only 25.9% and weakly supervised learning (45 registration pairs). Following [10], we decouple the semantic feature extraction and directly train an nnUNet model [19]. The best published VoxelMorph model that was trained with label-supervision and extended to a two-stream architecture reached 43.9% [27], the two top-ranked methods of the Learn2Reg challenge yield 65.7% (ConvexAdam [27]) and 67% (LapIRN [22]) respectively. Directly employing instance optimisation with 25 iterations on the nnUNet features achieves 62.9%. We use 2048 keypoints that are sampled inversely proportional to the predicted label maps and employ two warps (and inverse consistency for the first of them). Our model substantially outperforms VoxelMorph with 52.3% and sets a new state-of-the-art performance after instance optimisation reaching 69.6% with a total run time of less than a second.

4 Discussion and Conclusions

Our results demonstrate that contrary to previous belief, a simple single-scale U-Net architecture can provide large deformation estimation that is robust enough to reach high accuracy with a subsequent instance optimisation. The key insight of our work is the importance to predict a discretised heatmap to alleviate the problematic direct regression and use strong self-supervision either using automatic keypoint correspondences or a nonlocal multichannel loss together with a straightforward instance optimisation. Our work is related to mlVIRN [15], which also incorporates a keypoint loss for lung registration in addition to lobe segmentations, but has to be trained with several hundreds of paired CTs and did not report TRE values for DIRlab-COPD. Our network can be trained within 17 min on a single RTX A4000 requiring less than 2 GB of VRAM, indicating the improved training efficiency with fewer scans when using heatmaps. GraphRegNet [9] is similar in that it also employs heatmaps (integral regression) but more explicitly by defining the exact same discretised displacement grid beforehand and computing an SSD cost tensor based on hand-crafted features as input. While it outperforms our method with a TRE of 1.34 mm it appears to be more tailored towards the specific task and might not be easily extendable to end-to-end feature learning or abdominal registration.

References

1. 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)
2. Castillo, R., et al.: A reference dataset for deformable image registration spatial accuracy evaluation using the copdgene study archive. Phys. Med. Biol. **58**(9), 2861 (2013)
3. Castillo, R., et al.: A framework for evaluation of deformable image registration spatial accuracy using large landmark point sets. Phys. Med. Biol. **54**(7), 1849 (2009)
4. Eppenhof, K.A., Lafarge, M.W., Veta, M., Pluim, J.P.: Progressively trained convolutional neural networks for deformable image registration. IEEE Trans. Med. Imaging **39**(5), 1594–1604 (2019)
5. Estienne, T., et al.: MICS: multi-steps, inverse consistency and symmetric deep learning registration network (2021)
6. Falk, T., et al.: U-Net: deep learning for cell counting, detection, and morphometry. Nat. Methods **16**(1), 67–70 (2019)
7. Haber, E., Modersitzki, J.: Intensity gradient based registration and fusion of multi-modal images. Methods Inf. Med. **46**(03), 292–299 (2007)
8. Hansen, L., Dittmer, D., Heinrich, M.P.: Learning deformable point set registration with regularized dynamic graph CNNs for large lung motion in COPD patients. In: Zhang, D., Zhou, L., Jie, B., Liu, M. (eds.) GLMI 2019. LNCS, vol. 11849, pp. 53–61. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-35817-4_7
9. Hansen, L., Heinrich, M.P.: GraphregNet: deep graph regularisation networks on sparse keypoints for dense registration of 3d lung CTS. IEEE Trans. Med. Imaging **40**(9), 2246–2257 (2021)
10. Hansen, L., Heinrich, M.P.: Revisiting iterative highly efficient optimisation schemes in medical image registration. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12904, pp. 203–212. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87202-1_20
11. 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

12. Heinrich, M.P., Handels, H., Simpson, I.J.A.: Estimating large lung motion in COPD patients by symmetric regularised correspondence fields. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9350, pp. 338–345. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24571-3_41
13. Heinrich, M.P., Hansen, L.: Highly accurate and memory efficient unsupervised learning-based discrete CT registration using 2.5D displacement search. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12263, pp. 190–200. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59716-0_19
14. Heinrich, M.P., et al.: Mind: modality independent neighbourhood descriptor for multi-modal deformable registration. Med. Image Anal. **16**(7), 1423–1435 (2012)
15. Hering, A., Häger, S., Moltz, J., Lessmann, N., Heldmann, S., van Ginneken, B.: CNN-based lung CT registration with multiple anatomical constraints. Med. Image Anal., 102139 (2021)
16. Hering, A., et al.: Learn2Reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning (2021)
17. Hering, A., Murphy, K., van Ginneken, B.: Learn2Reg challenge: CT lung registration - training data, May 2020. https://doi.org/10.5281/zenodo.3835682
18. Hu, X., Kang, M., Huang, W., Scott, M.R., Wiest, R., Reyes, M.: Dual-stream pyramid registration network. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 382–390. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_43
19. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: NNU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)
20. Liu, X., Qi, C.R., Guibas, L.J.: FlowNet3d: learning scene flow in 3d point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 529–537 (2019)
21. Mok, T.C.W., Chung, A.C.S.: Conditional deformable image registration with convolutional neural network. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12904, pp. 35–45. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87202-1_4
22. 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
23. Murphy, K., et al.: Evaluation of registration methods on thoracic CT: the empire10 challenge. IEEE Trans. Med. Imaging **30**(11), 1901–1920 (2011)
24. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 483–499. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_29
25. Ou, Y., Sotiras, A., Paragios, N., Davatzikos, C.: DRAMMS: deformable registration via attribute matching and mutual-saliency weighting. Med. Image Anal. **15**(4), 622–639 (2011)
26. Sang, Y., Ruan, D.: Scale-adaptive deep network for deformable image registration. Med. Phys. **48**(7), 3815–3826 (2021)
27. Siebert, H., Hansen, L., Heinrich, M.P.: Fast 3d registration with accurate optimisation and little learning for learn2reg 2021 (2021)
28. Sotiras, A., Davatzikos, C., Paragios, N.: Deformable medical image registration: a survey. IEEE Trans. Med. Imaging **32**(7), 1153–1190 (2013)

29. Teed, Z., Deng, J.: RAFT: recurrent all-pairs field transforms for optical flow. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12347, pp. 402–419. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58536-5_24
30. de Vos, B.D., et al.: A deep learning framework for unsupervised affine and deformable image registration. Med. Image Anal. **52**, 128–143 (2019)
31. Xu, Z., et al.: Evaluation of six registration methods for the human abdomen on clinically acquired CT. IEEE Trans. Biomed. Eng. **63**(8), 1563–1572 (2016)
32. Zhao, S., Lau, T., Luo, J., Eric, I., Chang, C., Xu, Y.: Unsupervised 3d end-to-end medical image registration with volume tweening network. IEEE J. Biomed. Health Inform. **24**(5), 1394–1404 (2019)

Unsupervised Learning of Diffeomorphic Image Registration via TransMorph

Junyu Chen(✉), Eric C. Frey, and Yong Du

Russell H. Morgan Department of Radiology and Radiological Science, Johns Hopkins Medical Institutes, Baltimore, MD, USA
jchen245@jhmi.edu

Abstract. In this work, we propose a learning-based framework for unsupervised and end-to-end learning of diffeomorphic image registration. Specifically, the proposed network learns to produce and integrate time-dependent velocity fields in an LDDMM setting. The proposed method guarantees a diffeomorphic transformation and allows the transformation to be easily and accurately inverted. We also showed that, without explicitly imposing a diffeomorphism, the proposed network can provide a significant performance gain while preserving the spatial smoothness in the deformation. The proposed method outperforms the state-of-the-art registration methods on two widely used publicly available datasets, indicating its effectiveness for image registration. The source code of this work is available at: https://bit.ly/3EtYUFN.

Keywords: Image registration · Transformer · Deep neural networks

1 Introduction

Deformable image registration functions by establishing the spatial correspondence between the moving and the fixed images. Traditionally, image registration has been accomplished by optimizing a pair-wise objective function iteratively [3,5,9,18]. Over the last decade, deep learning has emerged as a major area of research in the field of medical image analysis, including registration [4,7,8,12,15,16,20]. Learning-based registration models optimize a global functional for a dataset during training, thereby obviating the time-consuming and computationally expensive per-image optimization during inference.

Diffeomorphic image registration is appealing in many medical imaging applications, owing to its properties like topology preservation and transformation invertibility. A diffeomorphic transformation can be achieved via the time integration of sufficiently smooth time-stationary [1,2,11] or time-dependent velocity fields [3,5]. Almost all existing *end-to-end* learning-based registration models adopt stationary velocity fields because of their ease of implementation and relatively low computational cost [8,15,16]. In this work, however, we demonstrate how time-dependent velocity fields can be efficiently incorporated into an *end-to-end* deep neural network framework, which results in diffeomorphisms (an illustrative example is shown in Fig. 1) and improved registration performance.

A. Hering et al. (Eds.): WBIR 2022, LNCS 13386, pp. 96–102, 2022.
https://doi.org/10.1007/978-3-031-11203-4_11

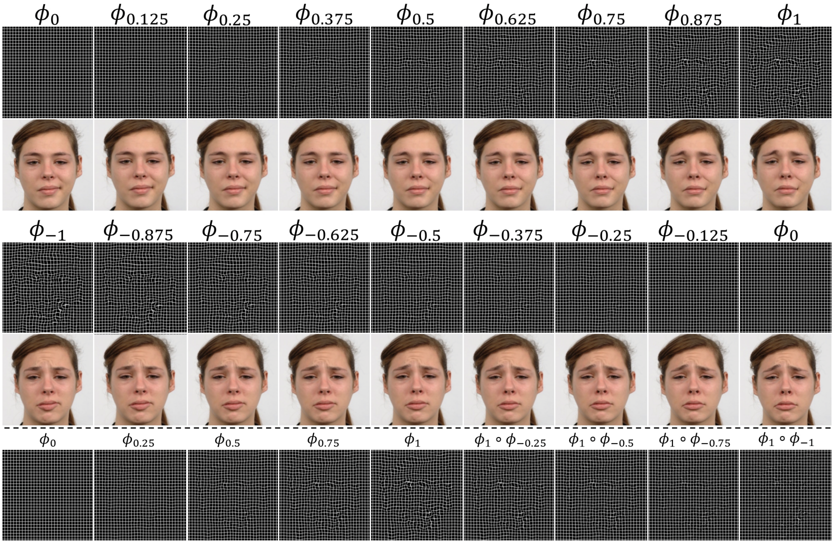


Fig. 1. Inversion and composition of the deformation fields using the proposed method. A neural network learns to generate time-dependent velocity fields for 8 time-steps.

2 Background on LDDMM

In the LDDMM setting [5], the transformation ϕ_t is computed as the flow of a time-dependent velocity field v_t, specified by the ODE: $\frac{d\phi}{dt} = v_t(\phi_t)$ with $t \in [0, 1]$. The final transformation at $t = 1$ is gained by integrating the velocity fields in time: $\phi_1 = \phi_0 + \int_0^1 v_t(\phi_t)dt$ with $\phi_0 = Id$. Then, the optimal transformation is formulated as a variational problem of the form:

$$v^* = \arg\min_v \left(\lambda \int_0^1 \|v_t\|_V^2 dt + \|I_0 \circ \phi_1 - I_1\|_{L^2}^2 \right), \tag{1}$$

where $\|\cdot\|_{L^2}$ denotes the standard L_2-norm, $\|f\|_V = \|Lf\|_{L^2}$ and L is a differential operator of the type $(-\alpha\varDelta + \gamma)^\beta Id$ with $\beta > 1.5$, and I_0 and I_1 are the moving and fixed images, respectively. With sufficiently smooth v, a dffieomorphism is guaranteed in this setting.

3 Methods

In this work, a neural network was used to generate velocity fields with a predetermined discretized number of time-steps, specified by N (as shown in Fig. 2). Then, the field integration layer integrates the generated velocity fields to form the transformation at the end-point, i.e., $\phi_1 \approx Id + \sum_{t=1}^N v_t \circ \phi_t$, and the inverse transformation ϕ_{-1} is computed as $Id - \sum_{t=1}^N v_t \circ \phi_t$. The proposed network may be trained self-supervisedly, end-to-end, using moving and fixed image pairs. We

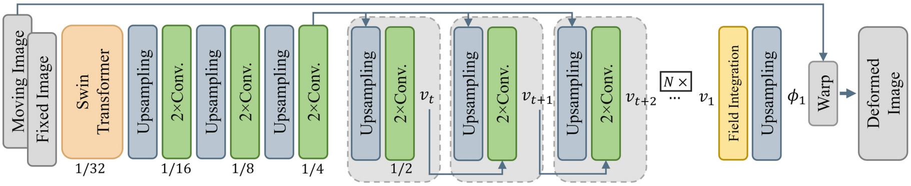


Fig. 2. Network architecture. The network integrates N time-steps of velocity fields to form a final deformation field. Note that skip connections and activation functions were omitted for visualization.

chose our previously developed `TransMorph` [6] (denoted as TM) as the base network since it showed state-of-the-art performance on several datasets. However, we underline that the proposed method is not architecture-specific and can readily be integrated into any architecture. The loss function was derived from Eq. 1 with an additional term to account the available label map information:

$$\mathcal{L}(v, I_0, I_1) = \sum_t \|Lv_t\|^2_{L^2} + \|I_0 \circ \phi_1 - I_1\|^2_{L^2} + \frac{1}{M}\sum_m \|S^m_0 \circ \phi_1 - S^m_1\|^2_{L^2}, \tag{2}$$

where S_0 and S_1 denote the M-channel label maps of the moving and fixed images, respectively, where each channel corresponds to the label map of an anatomical structure. We denote the model trained using this loss function as TM-TVF$_{LDDMM}$.

As a consequence of imposing a diffeomorphic transformation, excessive regularization may lead to a suboptimal registration accuracy measured by image similarity or segmentation overlap. Here, we demonstrate that by integrating time-dependent velocity fields, we could implicitly enforce transformation smoothness and improve performance without explicitly imposing a diffeomorphism. In this setting, we used a diffusion regularizer to regularize *only* the velocity field at the end-point:

$$\mathcal{L}(v, I_0, I_1) = \|\nabla v_1\|^2_{L^2} + NCC(I_0 \circ \phi_1, I_1) + Dice(S_0 \circ \phi_1, S_1), \tag{3}$$

where ∇v is the spatial gradient operator applied to v, $NCC(\cdot)$ denotes normalized cross-correlation, and $Dice(\cdot)$ denotes Dice loss. We denote the model trained using this loss function as TM-TVF.

4 Experiments and Results

We validated the proposed method using two publicly available datasets, one in 2D and one in 3D. The 2D dataset is the Radboud Faces Database (RaFD) [13], and it comprises eight distinct facial expression images for each of 67 subjects. We randomly divided the subjects into 53, 7, and 7 subjects, and used face images of subjects glancing in the direction of the camera. A total of 2968, 392, and 392 image pairs were used for training, validation, and testing. The images were cropped then resized into 256 × 256. The 3D dataset is the OASIS dataset

Table 1. SSIM [19] and FSIM [21] comparisons between the proposed method and the others on the RaFD dataset.

	VM-2 [4]	VM-diff [8]	CycleMorph [12]	TM [6]	TM-TVF$_{LDDMM}$	TM-TVF
SSIM↑	0.858 ± 0.038	0.805 ± 0.044	0.875 ± 0.038	0.899 ± 0.035	0.829 ± 0.049	$\mathbf{0.910 \pm 0.028}$
FSIM↑	0.669 ± 0.039	0.613 ± 0.041	0.687 ± 0.042	0.716 ± 0.043	0.620 ± 0.053	$\mathbf{0.734 \pm 0.033}$
% of $\lvert J_\phi \rvert \leq 0$ ↓	0.798 ± 0.812	<**0.001**	0.092 ± 0.163	0.190 ± 0.194	<**0.001**	0.062 ± 0.107
SDlogJ↓	0.086 ± 0.022	0.051 ± 0.011	0.059 ± 0.014	0.065 ± 0.016	$\mathbf{0.046 \pm 0.010}$	0.057 ± 0.013

Table 2. Validation and test results for the OASIS dataset from the 2021 Learn2Reg challenge [10]. The validation results came from the challenge's leaderboard, whereas the test results came directly from the challenge's organizers.

	Validation			Test		
	Dice↑	SDlogJ↓	HdDist95↓	Dice↑	SDlogJ↓	HdDist95↓
ConvexAdam [17]	0.846 ± 0.016	$\mathbf{0.067 \pm 0.005}$	1.500 ± 0.304	0.81	**0.07**	1.63
LapIRN [16]	0.861 ± 0.015	0.072 ± 0.007	1.514 ± 0.337	0.82	**0.07**	1.67
TM [6]	0.862 ± 0.014	0.128 ± 0.021	1.431 ± 0.282	0.820	0.124	1.656
TM-TVF$_{LDDMM}$	0.833 ± 0.016	0.090 ± 0.005	1.630 ± 0.353	–	–	–
TM-TVF	$\mathbf{0.869 \pm 0.014}$	0.094 ± 0.018	$\mathbf{1.396 \pm 0.297}$	**0.824**	0.090	**1.633**

[14] obtained from the 2021 Learn2Reg challenge [10]. This dataset comprises a total of 451 brain T2 MRI images, with 394, 19, and 38 images being used for training, validation, and testing, respectively. We trained the proposed method for 500 epochs using a learning rate of $1e^{-4}$. The number of time-steps, N, was empirically set to 8. We set $\alpha = 0.01$, $\gamma = 0.01$, and $\beta = 2$ for RaFD dataset, and $\alpha = 0.01$, $\gamma = 0.001$, and $\beta = 2$ for OASIS dataset. Note that due to the absence of segmentation in the RaFD dataset, the segmentation losses in Eqs. 1 and 2 were omitted.

Table 1 and 2 show quantitative results of the proposed models on the RaFD and OASIS datasets. On both datasets, the proposed TM-TVF yielded the highest performance against all other methods, including the first-ranking method (LapIRN [16]) from the Learn2Reg challenge. Specifically, TM-TVF outperformed its base network TM in image similarity and segmentation overlap on the two datasets, with p values < 0.0001 from paired t-tests. Although, a diffeomorphism was not explicitly guaranteed in TM-TVF, it still produced much smoother transformations than TM and VM measured by SDlogJ and the percentage of non-positive Jacobian determinant. On the other hand, although TM-TVF$_{LDDMM}$ guarantees a diffeomorphic transformation (as shown in Fig. 1, 3, and 4), it results in relatively poor registration performance, which is most likely owing to the excessive regularization imposed on the transformation.

5 Conclusion

In conclusion, we have proposed a learning-based framework for learning to generate time-dependent velocity fields in the LDDMM setting. The quantitative results show that the framework outperformed state-of-the-art registration models, indicating the effectiveness of the proposed method. Moreover, the proposed method is not architecture-specific and may be easily incorporated to improve registration performance in any network architecture.

Acknowledgment. This work was supported by a grant from the National Cancer Institute, U01-CA140204.

Appendix A. Additional Qualitative Results

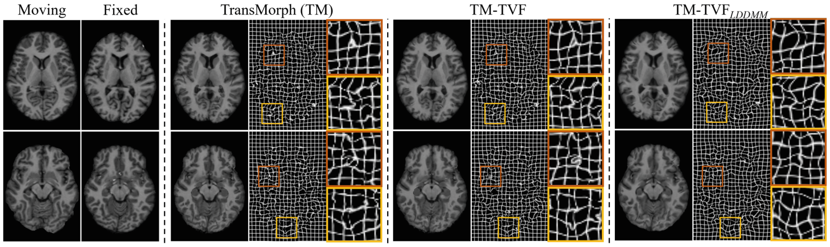


Fig. 3. Qualitative comparisons of the deformation field smoothness. TM yielded a deformation field with noticeable folded voxels, but TM-TVF generated a smoother field with state-of-the-art registration accuracy (as seen in Tables 1 and 2). TM-TVF$_{LDDMM}$ generated a highly regularized deformation field with nearly no visible folded voxels.

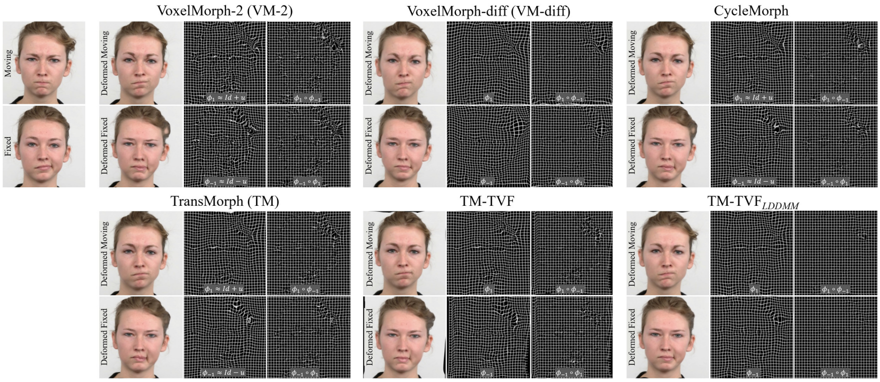


Fig. 4. Qualitative comparisons of facial expression registration. TM-TVF$_{LDDMM}$ produced a smooth and invertible transformation, but all other transformations were not. Additionally, TM-TVF yielded the best qualitative results for both forward and backward registration. Note that the transformation inversions for VM-2, CycleMorph, and TM were approximated using $Id - u$, where u denotes the displacement field.

References

1. Arsigny, V., Commowick, O., Pennec, X., Ayache, N.: A log-euclidean framework for statistics on diffeomorphisms. In: Larsen, R., Nielsen, M., Sporring, J. (eds.) MICCAI 2006. LNCS, vol. 4190, pp. 924–931. Springer, Heidelberg (2006). https://doi.org/10.1007/11866565_113
2. Ashburner, J.: A fast diffeomorphic image registration algorithm. Neuroimage **38**(1), 95–113 (2007)
3. Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C.: Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Med. Image Anal. **12**(1), 26–41 (2008)
4. 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)
5. Beg, M.F., Miller, M.I., Trouvé, A., Younes, L.: Computing large deformation metric mappings via geodesic flows of diffeomorphisms. Int. J. Comput. Vision **61**(2), 139–157 (2005)
6. Chen, J., Frey, E.C., He, Y., Segars, W.P., Li, Y., Du, Y.: Transmorph: transformer for unsupervised medical image registration (2021). https://arxiv.org/abs/2111.10480
7. Chen, J., He, Y., Frey, E.C., Li, Y., Du, Y.: ViT-V-Net: vision transformer for unsupervised volumetric medical image registration. In: Medical Imaging with Deep Learning (2021)
8. Dalca, A.V., Balakrishnan, G., Guttag, J., Sabuncu, M.R.: Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. Med. Image Anal. **57**, 226–236 (2019)
9. Heinrich, M.P., Jenkinson, M., Brady, M., Schnabel, J.A.: MRF-based deformable registration and ventilation estimation of lung CT. IEEE Trans. Med. Imaging **32**(7), 1239–1248 (2013)
10. Hering, A., et al.: Learn2reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning. arXiv preprint arXiv:2112.04489 (2021)
11. Hernandez, M., Bossa, M.N., Olmos, S.: Registration of anatomical images using paths of diffeomorphisms parameterized with stationary vector field flows. Int. J. Comput. Vision **85**(3), 291–306 (2009)
12. Kim, B., et al.: CycleMorph: cycle consistent unsupervised deformable image registration. Med. Image Anal. **71**, 102036 (2021)
13. Langner, O., et al.: Presentation and validation of the radboud faces database. Cogn. Emot. **24**(8), 1377–1388 (2010)
14. Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C., Buckner, R.L.: Open access series of imaging studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. J. Cogn. Neurosci. **19**(9), 1498–1507 (2007)
15. Mok, T.C., Chung, A.: Fast symmetric diffeomorphic image registration with convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4644–4653 (2020)
16. Mok, T.C.W., Chung, A.C.S.: Conditional deformable image registration with convolutional neural network. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12904, pp. 35–45. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87202-1_4

17. Siebert, H., Hansen, L., Heinrich, M.P.: Fast 3d registration with accurate optimisation and little learning for learn2reg 2021. arXiv preprint arXiv:2112.03053 (2021)
18. Vercauteren, T., Pennec, X., Perchant, A., Ayache, N.: Diffeomorphic demons: efficient non-parametric image registration. Neuroimage **45**(1), S61–S72 (2009)
19. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
20. Yang, X., Kwitt, R., Styner, M., Niethammer, M.: Quicksilver: fast predictive image registration-a deep learning approach. Neuroimage **158**, 378–396 (2017)
21. Zhang, L., Zhang, L., Mou, X., Zhang, D.: FSIM: a feature similarity index for image quality assessment. IEEE Trans. Image Process. **20**(8), 2378–2386 (2011)

SuperWarp: Supervised Learning and Warping on U-Net for Invariant Subvoxel-Precise Registration

Sean I. Young[1,2](✉), Yaël Balbastre[1,2], Adrian V. Dalca[1,2], William M. Wells[1,2], Juan Eugenio Iglesias[1,2], and Bruce Fischl[1,2]

[1] MGH/HST Martinos Center for Biomedical Imaging, Boston, USA
{siyoung,adalca}@mit.edu, ybalbastre@mgh.harvard.edu
[2] Massachusetts Institute of Technology, Cambridge, USA

Abstract. In recent years, learning-based image registration methods have gradually moved away from direct supervision with target warps to self-supervision using segmentations, producing promising results across several benchmarks. In this paper, we argue that the relative failure of supervised registration approaches can in part be blamed on the use of regular U-Nets, which are jointly tasked with feature extraction, feature matching, and estimation of deformation. We introduce one simple but crucial modification to the U-Net that disentangles feature extraction and matching from deformation prediction, allowing the U-Net to warp the features, across levels, as the deformation field is evolved. With this modification, direct supervision using target warps begins to outperform self-supervision approaches that require segmentations, presenting new directions for registration when images do not have segmentations. We hope that our findings in this preliminary workshop paper will re-ignite research interest in supervised image registration techniques. Our code is publicly available from https://github.com/balbasty/superwarp.

Keywords: Image registration · Optical flow · Supervised learning

1 Introduction

In recent years, fully convolutional networks (FCNs) have become a universal framework for tackling an array of problems in medical imaging, ranging from image denoising and super-resolution [1, 2] to semantic segmentation [3–5], and from style transfer [6, 7] to image registration. Among these, image registration methods have benefitted immensely from FCNs, allowing methods to transition from an optimization-based paradigm to a learning-based one and to accelerate the alignment of images with different contrasts or modalities, for example.

An overwhelming majority of recent image registration networks [8–11] are trained unsupervised, in the sense that ground-truth deformation fields are not required in the supervision of these networks. Instead, a surrogate photometric loss is used to maximize the similarity between the fixed image and the moving one—warped by the predicted

A. Hering et al. (Eds.): WBIR 2022, LNCS 13386, pp. 103–115, 2022.
https://doi.org/10.1007/978-3-031-11203-4_12

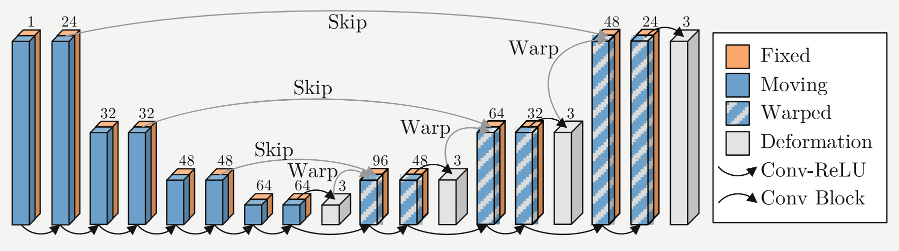


Fig. 1. The SuperWarp U-Net for image registration (first four levels shown). The fixed and moving images are concatenated along the batch axis and processed through the network. The two image features are reconcatenated along the channel axis at each level of the U-Net's upward path to be processed into a residual deformation, used to warp the moving features, and scaled and summed to produce the final deformation.

deformation field—in lieu of a loss that penalizes the differences between the predicted and ground-truth deformation fields. Since images typically contain large untextured regions as well as different contrasts and voxel intensities, merely minimizing differences in the fixed image and the moving one is insufficient to recover the ground truth deformation, even when a smoothness prior (or regularization) is imposed on the predicted deformation field. While supervised [12–14] and self-supervised approaches [9, 10]—based on segmentations, for example—produce excellent results, direct supervision using target warps is still desirable in many cases especially if the images do not have segmentations. However, supervised registration has not been as successful for many applications due to severe optimization difficulties faced—the network is jointly tasked with feature extraction and matching in addition to deformation estimation, which is not handled well by a fully convolutional network.

In this work, we will propose SuperWarp, a supervised learning approach to medical image registration. We first re-visit the classic optical flow equation of Horn and Schunck [15] to analyze its implications for supervised registration—the duality of intensity-invariant feature extraction and deformation estimation and the need for multi-scale warping. With such implications in mind, we make one simple but critical modification to the segmentation U-Net that repurposes it for subvoxel- (or subpixel-) accurate supervised image registration. With this modification, direct supervision using target warps outperforms self-supervised registration requiring segmentations. The network, shown in Fig. 1, is strikingly similar to a segmentation U-Net except for warping and deformation extraction layers, allowing U-Net to warp the features as the deformation field is evolved.

2 Related Work

SuperWarp is heavily inspired by previous work on optical flow estimation, the aim of which is to recover apparent motion from an image pair [15]. Expressed mathematically, however, optical flow estimation and image registration are in fact identical problems possibly except for the notion of regularity in each—an optical flow field for the former is typically assumed differentiable a.e. whereas a deformation field for the latter infinitely differentiable or diffeomorphic. This subtle distinction between the two problems does however disappear under the supervised learning paradigm since the type of regularity desired is reflected in the ground-truth optical flow (or deformation) fields of the training data.

2.1 Optical Flow Estimation

Here, we briefly recap development in classical and learning-based optical flow estimation methods—see e.g. [16] for a review. In their seminal work, Horn and Schunck [15] formulated optical flow estimation via a regularized optimization problem, noting that the problem is generally ill-posed in the absence of local smoothness priors. Several works extend the original Horn–Schunck model [15] using sub-quadratic regularization and data fidelity terms [17–21] that mitigate the deleterious effects of occlusions on flow estimation. Oriented regularization terms [21–25] regularize the flow only along the direction tangent to the image gradient while non-local terms [26–29] regularize flow even across disconnected pixels subject to similar motion. Median filtering of intermediate flows [23, 26] achieves similar effects to non-local regularity terms. Higher-order regularizers [28, 30] assign zero penalty to affine trends in the flow to encourage piecewise- linear flow predictions. Despite the advances, designing a regularizer is highly domain-specific, suggesting that it can be alleviated via supervised learning.

Orthogonally to the choice of regularizers, multi-scale schemes [31–34] have been used to estimate larger flows. Descriptor matching [31, 32] introduces an extra data fidelity term that penalizes misalignment of scale-invariant features (e.g. SIFT), overcoming the deterioration of the conventional data fidelity term at large scales due to the loss of small image structures. Since the optical flow equation no longer holds in the presence of a global brightness change, several authors propose to attenuate the brightness component of the images as a first step using high-pass filters [18, 24, 35], structure-texture decomposition [27, 36] or color space transforms [24]. Thus, in traditional approaches, both multi-scale processing and brightness-invariant transforms require us to handcraft suitable pre-processing filters, which can be highly time-consuming owing to the image-dependent nature of such filters. As we will see, the U-Net architecture used in the SuperWarp obviates the need to handcraft such filters, allowing the U-Net to learn them directly from the training data, end-to-end, to enable brightness- invariant image registration with exceptional generalization ability.

Fischer et al. [37] formulate optical flow estimation as a supervised learning problem. They train a U-Net model to output the optical flow field directly for a pair of input images, supervising the training using the ground-truth optical flow field as the target. Later works extend [37], cascading multiple instances of the network with warping [38], introducing a warping layer [39] or using a fixed image pyramid [40] to improve the accuracy of flow prediction [38, 39] as well as reduce the model size. Some authors propose to tackle optical flow estimation as an unsupervised learning task [41, 42], using a photometric loss to penalize the intensity differences across the fixed and moved images. Recent extensions in this unsupervised direction include occlusion-robust losses [43, 44] based on forward-backward consistency, and self-supervision losses [42, 45]. These are also the building blocks of unsupervised image registration methods [9–11].

3 Mathematical Framework

3.1 Optical Flow Estimation and Duality Principle

Under a sufficiently high temporal sampling rate, we can relate the intensities of a successive pair of three-dimensional images $(\mathbf{f}_0, \mathbf{f}_1)$ to components $(\mathbf{u}, \mathbf{v}, \mathbf{w})$ of the displacement between the two images using the optical flow equation

$$(\partial \mathbf{f}_1/\partial x) \cdot \mathbf{u} + (\partial \mathbf{f}_1/\partial y) \cdot \mathbf{v} + (\partial \mathbf{f}_1/\partial z) \cdot \mathbf{w} = \mathbf{f}_0 - \mathbf{f}_1 \tag{1}$$

[15], where $(\partial/\partial x, \partial/\partial y, \partial/\partial z)$ denotes the 3D spatial gradient operator. PDE (1) can also be seen as a linearization of the small deformation model in image registration [46]. Since (1) involves three unknowns for every equation, finding $(\mathbf{u}, \mathbf{v}, \mathbf{w})$ given $(\mathbf{f}_0, \mathbf{f}_1)$ is an ill-posed inverse problem. Smoothness assumptions are therefore made in optimization-based flow estimation [17–21] to render the inverse problem well-posed again similar to image registration [9–11].

A global change in the brightness or contrast across the image pair $(\mathbf{f}_0, \mathbf{f}_1)$ introduces an additive bias in the right-hand side of (1) such that the equation no longer holds. Compensating for this change in pre-processing would require knowledge of the displacement field $(\mathbf{u}, \mathbf{v}, \mathbf{w})$ that we seek in the first place. A similar issue is often met in medical image registration, with different imaging modalities across $\mathbf{f}_0$ and $\mathbf{f}_1$ injecting additive and multiplicative biases in (1). If however we knew the ground-truth displacement $(\mathbf{u}, \mathbf{v}, \mathbf{w})$, harmonizing $(\mathbf{f}_0, \mathbf{f}_1)$ in a normalized intensity space is readily achieved via (1). Conversely, given a harmonized image pair, the displacement field can be recovered using (1).

Image segmentation [47] is the ultimate form of image harmonization, since it removes brightness and contrast from images altogether and turns them into piecewise smooth (constant) signals by construction. This suggests that the use segmentation maps to supervise registration [9, 10] can be beneficial. However, many types of images do not have segmentations available or lack the notion of segmentation altogether, e.g. fMRI activations, so supervision using the ground-truth warps instead can be an expedient way of learning to register.

In practice, images $(\mathbf{f}_0, \mathbf{f}_1)$ are acquired at a low temporal sampling rate so (1) holds only over regions where both image intensities are linear functions of their spatial coordinates [15]. Equivalently, (1) holds in the general case only if the magnitudes of the components $(\mathbf{u}, \mathbf{v}, \mathbf{w})$ are less than one voxel. Since this can pose a major limitation for practical applications, multi-scale processing is used to linearize the images at gradually smaller scales, with the displacement field estimated at the larger scale used to initialize the residual flow estimation at the smaller scale. Linearizing images at larger scales, however, results in the loss of small structures due to the smoothing filters. Handcrafting filters that have an optimum tradeoff between linearization and preservation of image features at every scale is image-dependent and can be time-consuming, implying that learning such filters end-to-end can be beneficial for generalization ability.

3.2 Supervised Learning and Multi-scale Warping

SuperWarp exploits the duality principle (1) to supervise an image registration network equipped with multi-scale warping to estimate large deformations. We train a U-Net model on pairs of images with different intensities related via our smoothly synthesized ground-truth deformation fields. The downward path of the U-Net model first extracts intensity-invariant features from the two images separately. The upward path then extracts from the feature pair a deformation that minimizes the differences with respect to the ground-truth target.

SuperWarp makes one important modification to the registration U-Net for large displacement estimation. At each level of the network's upward path, the features of the moving image are first warped using the deformation field from the previous level, such that only the residual deformation, less than a voxel in magnitude, need be extracted at the current level. Processing the two images jointly as a single multi-channel image through U-Net, as done in [9, 10], would entangle the features of the fixed and the moving images, so that warping only the features of the moving one post hoc is not feasible. Instead, we process the two images as a batch with the image pair interacting only during deformation extraction, where the two image features are reconcatenated along the channels axis and processed into deformation field using a convolution block.

Note from the left and the right-hand sides of (1) that it is $(\mathbf{f}_1, \mathbf{f}_0 - \mathbf{f}_1)$, not $(\mathbf{f}_1, \mathbf{f}_0)$, which needs to be processed for displacement estimation. This suggests that feeding the features of $\mathbf{f}_1$ and pre-computed feature differences between $\mathbf{f}_0$ and (warped) $\mathbf{f}_1$ into deformation blocks can yield a saving of one convolution layer per block, which is substantial given that these blocks typically have no more than three convolution layers in total. In practice, we reparameterize the input further to the features of $(\mathbf{f}_0 + \mathbf{f}_1, \mathbf{f}_0 - \mathbf{f}_1)$ to help the extraction blocks average the spatial derivatives of the features across the two images, similarly to the practice in optimization-based approaches [16]. This reparameterization can be seen as a Hadamard transform [48] across the two image feature sets.

Table 1. Parameter ranges and probabilities used for random spatial transformation and intensity augmentation of the image pair. Applied separately to each image in the pair.

	Spatial Transformation					**Intensity Augmentation**			
	Translate	Scale	Rotate	Shear	Elastic	Noise Std	Multiply	Contrast	Gamma
Range	±12	[0.75,1.25]	±30°	±0.012	±4 (256^2)	[0,0.05]	[0.75,1.25]	[0.75,1.25]	[0.70,1.50]
Prob.	1.0	1.0	1.0	1.0	1.0	0.5	0.5	0.5	0.5

3.3 Deep Supervision, Data Augmentation and Training

Following the approach of deep supervision for semantic segmentation [49], we supervise the deformation block at each level of the U-Net's upward path with a displacement target. We use the MSE loss between the predicted $(\mathbf{u}, \mathbf{v}, \mathbf{w})$ and the target $(\mathbf{p}, \mathbf{q}, \mathbf{r})$ to minimize $E(\mathbf{u}, \mathbf{v}, \mathbf{w}) = \|(\mathbf{u}, \mathbf{v}, \mathbf{w}) - (\mathbf{p}, \mathbf{q}, \mathbf{r})\|_2^2$. The loss is summed across levels without weighting to produce the final training loss. The deformation block at each level is supervised using the target ground-truth field down-sampled to the spatial dimensions of its predictions. For evaluation, more forgiving EPE loss $E_{\text{EPE}}(\mathbf{u}, \mathbf{v}, \mathbf{w}) = \|(\mathbf{u}, \mathbf{v}, \mathbf{w}) - (\mathbf{p}, \mathbf{q}, \mathbf{r})\|_{2,1}$ is used instead.

To generate training pairs of images with their corresponding ground-truth deformation targets, we sample an image $\mathbf{f}$ from the training set and synthesize two different smooth displacements $(\mathbf{p}_0, \mathbf{q}_0, \mathbf{r}_0)$ and $(\mathbf{p}_1, \mathbf{q}_1, \mathbf{r}_1)$ that warp $\mathbf{f}$ and produce $\mathbf{f}_0$ and $\mathbf{f}_1$, respectively. The ground-truth displacement is given by

$$(\mathbf{p}, \mathbf{q}, \mathbf{r}) = (\mathbf{Id} + (\mathbf{p}_1, \mathbf{q}_1, \mathbf{r}_1))^{-1}(\mathbf{Id} + (\mathbf{p}_0, \mathbf{q}_0, \mathbf{r}_0)) - \mathbf{Id}, \tag{2}$$

in which the identity $\mathbf{Id}$ denotes the (vectorization) of the grid coordinates. To facilitate computation, we restrict $(\mathbf{p}_1, \mathbf{q}_1, \mathbf{r}_1)$ to affine fields so that the inverse coordinate mapping $(\cdot)^{-1}$ (2) can be computed by inverting a 4 × 4 matrix. We apply a small elastic deformation on $(\mathbf{p}_0, \mathbf{q}_0, \mathbf{r}_0)$ to approximate a higher-order (non-affine) component of the spatial distortion typically seen in MR scans. We then transform the voxel intensities of $\mathbf{f}_0$ and $\mathbf{f}_1$ using a standard augmentation pipeline (Gaussian noise, brightness multiplication, contrast augmentation, and gamma transform); see Table 1 for the hyperparameters of these transforms.

For training, we use a batch size of 1, which actually becomes 2 because the moving and fixed images are concatenated along the batch axis. The Adam [50] optimizer is used with an initial learning rate of 10^{-4}, linearly reduced to 10^{-6} across 200,000 iterations. We find it beneficial to initially train the network for 20,000 iterations on training examples with zero displacement and deformation but still with intensity augmentations to enable the network to learn to extract contrast-invariant features, then introducing deformations to train the network to predict deformations with brightness change across the image pair.

In Fig. 2, we plot validation Dice and end-point error curves of SuperWarp U-Net (ours) and a VoxelMorph-like U-Net baseline for the registration of MR brain scans. In the Dice-supervised case, we train the networks to minimize the regularized Dice loss between the segmentations of the fixed and moving images

$$E_{\text{Dice}}(\mathbf{u}, \mathbf{v}, \mathbf{w}) = D_{\text{Dice}}(\mathbf{f}_1 \circ (\mathbf{Id} + (\mathbf{u}, \mathbf{v}, \mathbf{w})), \mathbf{f}_0) + R(\mathbf{u}, \mathbf{v}, \mathbf{w}), \tag{3}$$

in which R penalizes the (squared) Laplacian of the components $\mathbf{u}, \mathbf{v}, \mathbf{w}$. In the MSE-supervised case, the same networks are trained to minimize the MSE in the predicted and target deformations. Regardless of the training objective, our SuperWarp U-Net outperforms the baseline U-Net and also trains significantly faster, requiring only 20 iterations to reach maximum accuracy in the case where the Dice loss is used. Moreover, SuperWarp U-Net trained using the MSE loss (no segmentations) outperforms the Dice baseline requiring segmentations.

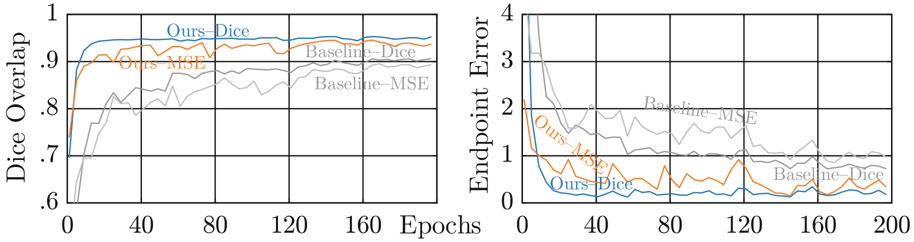


Fig. 2. Validation registration accuracy. In both self-supervised (MSE) and supervised (Dice) cases, the SuperWarp U-Net leads to better mean Dice and endpoint error than the baseline (similar to VoxelMorph) and trains faster, requiring only 40 epochs to reach the final accuracy, which are 0.954, 0.152 (Ours–Dice), 0.906, 0.711 (Baseline–Dice).

4 Experimental Evaluation

We validate our proposed SuperWarp approaches on two datasets—a set of 2D brain magnetic resonance (MR) scans, as well as the Flying Chairs [37] optical flow dataset widely used in computer vision. The brain image registration task allows us to benchmark the performance of SuperWarp against related work in medical image registration [9, 10] while Flying Chairs allows us to compare the SuperWarp U-Net with the state-of-the-art optical flow estimation networks. In addition to Dice scores between fixed and moved images, we also use the mean EPE to evaluate the accuracy of the displacements. All U-Nets have 7 levels of [24, 32, 48, 64, 96, 128, 192] features and two convolution layers at each level.

4.1 Invariant Registration of Brain MR Images

Here, we apply SuperWarp to deformable registration of 2D brain scans within a subject. Obviously, SuperWarp could be applied to the cross-subject setting too but the accuracy of predicted deformations is easier to assess in the within-subject case and facilitates comparisons with other methods. We use the whole brain dataset of [51] containing 40 T1-weighted brain MR scans, along with the corresponding segmentations produced using FreeSurfer [51]. For test, we use a collection of 500 T1-weighted brain MR scans curated from: OASIS, ABIDE-I and -II, ADHD, COBRE, GSP, MCIC, PPMI, and UK Bio. The scan pairs are generated as described in Sect. 3.3. We do not perform linear registration of the images as a preprocessing step in any of the methods since the displacements are rather small (Table 1) and this provides better insights into their behavior.

To show the improvement in the accuracy of the deformation field recovered using our methods, we plot statistics of the validation end-point error and Dice scores produced by all methods including the baseline—similar to VoxelMorph [9]—in Fig. 3. While our Dice scores are higher than those of the baseline only by 0.04, our end-point errors are more significantly reduced from the baselines (by 80%, on average across, foreground pixels). Figure 4 shows the displacements predicted by our method, comparing them with those from the baseline.

To better understand the sources of improvement between the baseline and our approach, we conduct an extensive set of ablation studies on SuperWarp as listed in Table 2. We see that the multi-scale loss used in [49] can actually hurt accuracy for this experiment. Training with the EPE loss produces a worse EPE than training with the MSE loss likely due to numerical instability at zero. The number of U-Net levels should also be high enough (seven) to cover the largest displacements (about ± 64) at the coarsest level of the U-Net.

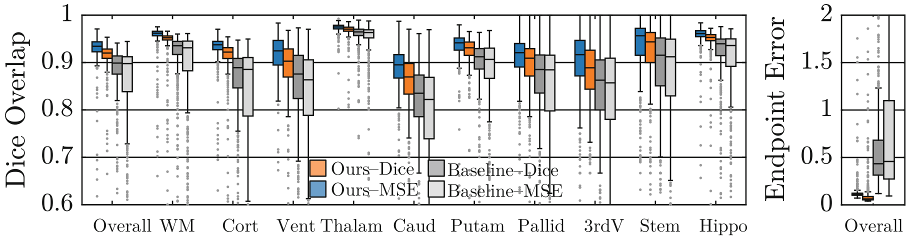


Fig. 3. Test Dice (left) and endpoint error (right) statistics on 10 structures across 500 T1w brain images. Regardless of the choice of the training loss function, the SuperWarp produces better Dice and endpoint error than the baseline (similar to VoxelMorph). Note that Ours–MSE does not need or use segmentation information.

4.2 Optical Flow Estimation

To further benchmark the network architecture used by the SuperWarp, we run additional experiments on the Flying Chairs optical flow dataset [37], popularly used by the computer vision community. To facilitate a fair comparison, we set our network and training hyperparameters very similarly to [39]: 7 U-Net levels for a total of 6.9M learnable parameters, 1M steps, EPE loss for training, multi-scale loss (but weight all scales equally) with the Adam optimizer. Table 3 lists the validation EPE of flow fields predicted using the SuperWarp and other well-performing models.

Both PWC-Net [39] and FlowNet-C [37] attribute their good performance to the use of the cost-volume layer but we find cost volumes to be unnecessary to achieve a good accuracy at least on this dataset. While SPY-Net [40] also uses a multi-scale warping strategy, it is based on a fixed image pyramid. This helps to bring down the number of trainable parameters but can also lead to a loss of image structure at coarser levels. FlowNet2 cascades multiple FlowNet models and warps in between, while the SuperWarp U-Net incorporates warps directly in the model, significantly reducing the model size with a comparable accuracy.

5 Discussion

In this paper we have shown that supervising an image registration network with a target warp can achieve state-of-the-art accuracy. Our approach outperforms previous supervised ones due to the multi-scale nature of our prediction, where deformations are composed across the upward path of the U-Net and applied to the features of the moving image. This way, each spatial scale receives a moving image as input that has

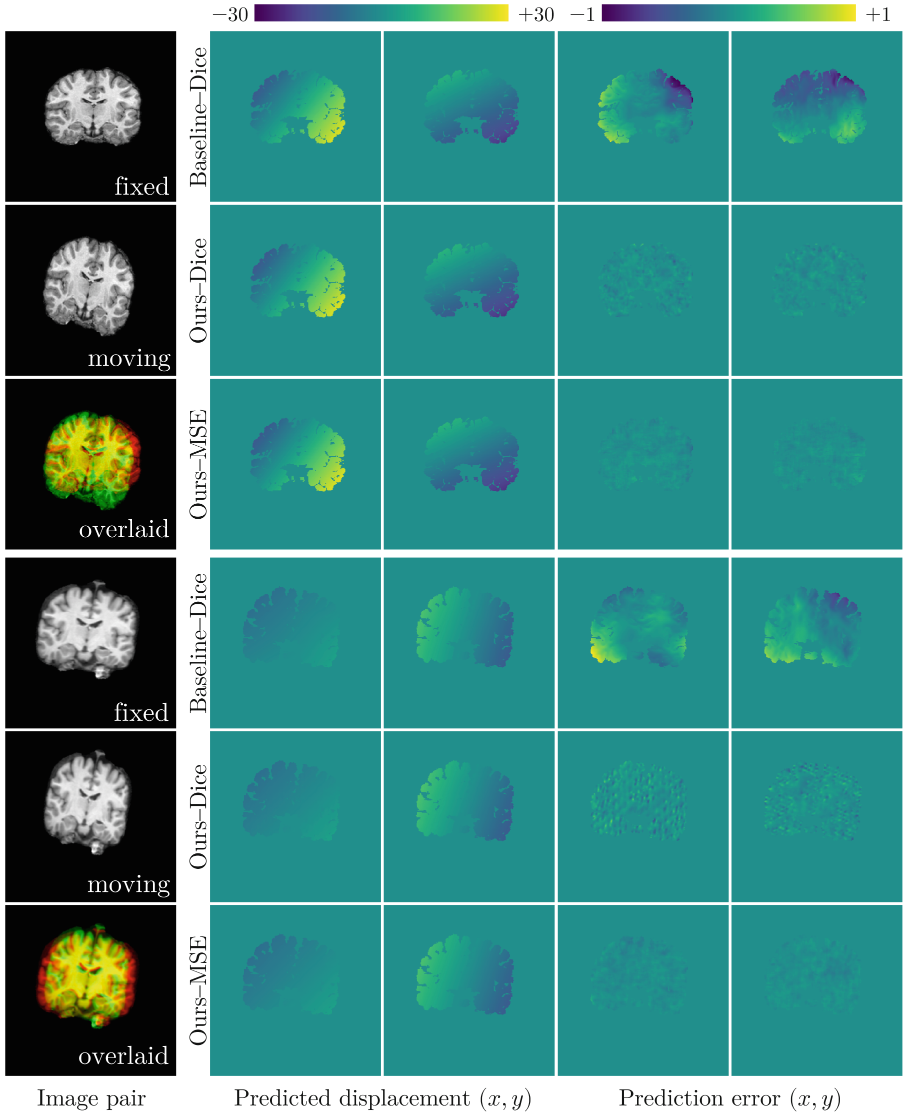


Fig. 4. Visualization of predicted displacement fields (test). Both Dice and MSE variants of the SuperWarp can produce highly accurate displacements (in the first example, 0.08 and 0.06 mm, respectively) whereas the baseline (similar to VoxelMorph) prediction has larger errors (0.41 mm). Images are 2D, 256 × 256, 1 mm isotropic. Cf. Fig. 3 (right).

Table 2. Ablation of network and training hyperparameters used and their influence on the best epoch validation accuracy. Default hyperparameter: (7, MSE, False, True).

	Number of levels		**Training loss function**			**Multi-scale loss**		**Multi-scale warp**	
	6	7	Dice	EPE	MSE	True	False	True	False
Dice	0.927	**0.947**	**0.954**	0.942	0.947	0.939	**0.947**	**0.947**	0.903
EPE	0.450	**0.122**	**0.103**	0.195	0.122	0.270	**0.122**	**0.122**	0.738

been warped by the composition of all larger spatial scales, ensuring that the optical flow condition holds for the deformation at that level. This recovers the accuracy of the deformation estimation that was likely lost in previous supervised techniques due to the lack of multi-scale warping.

Table 3. Mean EPE achieved by various network models on the Flying Chairs test set.

	PWC-Net	**SPY-Net**	**FlowNetS**	**FlowNetC**	**FlowNet2**	**Ours–EPE**
Parameters	8.75M	**1.20M**	32.1M	32.6M	64.2M	6.9M
EPE	2.00	2.63	2.71	2.19	**1.78**	1.82

While segmentation accuracy is itself of course important, we also point out that there are instances in which it is important to recover an exact deformation field. In these cases, using a segmentation loss leads to inaccuracies when there are too few segmentation classes to guide thedeformation estimation. We show that using the architecture we have described, we are able to recover an excellent prediction of a true underlying deformation field. Uses cases include distortion estimation and removal in MRI, such as those caused by inhomogeneities in the main magnetic field (B0) and image distortions induced by nonlinearities in the gradient coils used to encode spatial location.

5.1 Future Work

In this workshop paper, we have addressed only one type of invariance, namely invariance to intensity (or illumination) change across images. In the sequel, we plan to add contrast and distortion invariance to the network by training it on synthetic scans of various contrasts as done in [52] and applying the synthetic approach to distortions as well. Also, we plan to run a more comprehensive set of experiments on 3D MR images, showing the benefits of our approach in many clinical applications.

Acknowledgments. Support for this research provided in part by the BRAIN Initiative Cell Census Network grant U01MH117023, NIBIB (P41EB015896, 1R01EB023281, R01EB-006758, R21EB018907, R01EB019956, P41EB030006, P41EB028741), NIA (1R-56AG064027, 1R01AG064027, 5R01AG008122, R01AG016495, 1R01AG070988), NIMH (R01MH123195, R01MH121885, 1RF1MH123195), NINDS (R01NS05–25851, R21-NS072652, R01NS070963, R01NS083534, 5U01NS086625, 5U24NS-10059103, R01NS105820), ARUK (IRG2019A-003), and was made possible by resources from Shared Instrumentation Grants 1S10RR023401, 1S10RR019307, and 1S10-RR023043. Additional support was provided by the NIH Blueprint for Neuroscience Research (5U01-MH093765), part of the multi-institutional Human Connectome Project.

References

1. Zhang, K., Zuo, W., Chen, Y., Meng, D., Zhang, L.: Beyond a Gaussian denoiser: residual learning of deep CNN for image denoising. IEEE Trans. Image Process. **26**, 3142–3155 (2017)
2. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE Trans. Pattern Anal. Mach. Intell. **38**, 295–307 (2016)
3. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
4. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of CVPR, pp. 3431–3440 (2015)
5. Noh, H., Hong, S., Han, B.: Learning deconvolution network for semantic segmentation. In: Proceedings of CVPR (2015)
6. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of CVPR, pp. 2414–2423 (2016)
7. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 694–711. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_43
8. de Vos, B.D., Berendsen, F.F., Viergever, M.A., Staring, M., Išgum, I.: End-to-end unsupervised deformable image registration with a convolutional neural network. In: Cardoso, M.J., et al. (eds.) DLMIA/ML-CDS -2017. LNCS, vol. 10553, pp. 204–212. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67558-9_24
9. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: An unsupervised learning model for deformable medical image registration. In: Proceedings of CVPR (2018)
10. Dalca, A.V., Balakrishnan, G., Guttag, J., Sabuncu, M.R.: Unsupervised learning for fast probabilistic diffeomorphic registration. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11070, pp. 729–738. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00928-1_82
11. Mok, T.C.W., Chung, A.C.S.: Fast symmetric diffeomorphic image registration with convolutional neural networks. In: Proceedings of CVPR (2020)
12. Cao, X., et al.: Deformable image registration based on similarity-steered CNN regression. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10433, pp. 300–308. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66182-7_35
13. Rohé, M.-M., Datar, M., Heimann, T., Sermesant, M., Pennec, X.: SVF-Net: learning deformable image registration using shape matching. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10433, pp. 266–274. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66182-7_31
14. Yang, X., Kwitt, R., Styner, M., Niethammer, M.: Quicksilver: fast predictive image registration – a deep learning approach. Neuroimage **158**, 378–396 (2017)
15. Horn, B.K.P., Schunck, B.G.: Determining optical flow. Artif. Intell. **17**, 185–203 (1981)
16. Sun, D., Roth, S., Black, M.J.: A quantitative analysis of current practices in optical flow estimation and the principles behind them. Int. J. Comput. Vis. **106**, 115–137 (2014)

17. Black, M.J., Anandan, P.: The robust estimation of multiple motions: parametric and piecewise-smooth flow fields. Comput. Vis. Image Underst. **63**, 75–104 (1996)
18. Papenberg, N., Bruhn, A., Brox, T., Didas, S., Weickert, J.: Highly accurate optic flow computation with theoretically justified warping. Int. J. Comput. Vis. **67**, 141–158 (2006)
19. Roth, S., Lempitsky, V., Rother, C.: Discrete-continuous optimization for optical flow estimation. In: Cremers, D., Rosenhahn, B., Yuille, A.L., Schmidt, F.R. (eds.) Statistical and Geometrical Approaches to Visual Motion Analysis. LNCS, vol. 5604, pp. 1–22. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-03061-1_1
20. Zach, C., Pock, T., Bischof, H.: A duality based approach for realtime TV-L 1 optical flow. In: Hamprecht, F.A., Schnörr, C., Jähne, B. (eds.) Pattern Recognition, pp. 214–223. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-74936-3_22
21. Sun, D., Roth, S., Lewis, J.P., Black, M.J.: Learning Optical Flow. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008. LNCS, vol. 5304, pp. 83–97. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-88690-7_7
22. Nagel, H., Enkelmann, W.: An investigation of smoothness constraints for the estimation of displacement vector fields from image sequences. IEEE Trans. Pattern Anal. Mach. Intell. **8**, 565–593 (1986)
23. Wedel, A., Cremers, D., Pock, T., Bischof, H.: Structure- and motion-adaptive regularization for high accuracy optic flow. In: Proceedings of ICCV, pp. 1663–1668 (2009)
24. Zimmer, H., Bruhn, A., Weickert, J.: Optic flow in harmony. Int. J. Comput. Vis. **93**, 368–388 (2011)
25. Zimmer, H., et al.: Complementary optic flow. In: Cremers, D., Boykov, Y., Blake, A., Schmidt, F.R. (eds.) EMMCVPR 2009. LNCS, vol. 5681, pp. 207–220. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-03641-5_16
26. Sun, D., Roth, S., Black, M.J.: Secrets of optical flow estimation and their principles. In: Proceedings of CVPR, pp. 2432–2439 (2010)
27. Werlberger, M., Pock, T., Bischof, H.: Motion estimation with non-local total variation regularization. In: Proceedings of CVPR, pp. 2464–2471 (2010)
28. Ranftl, R., Bredies, K., Pock, T.: Non-local total generalized variation for optical flow estimation. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8689, pp. 439–454. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10590-1_29
29. Krähenbühl, P., Koltun, V.: Efficient nonlocal regularization for optical flow. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7572, pp. 356–369. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33718-5_26
30. Bredies, K., Kunisch, K., Pock, T.: Total generalized variation. SIAM J. Imaging Sci. **3**, 492–526 (2010)
31. Liu, C., Yuen, J., Torralba, A.: Sift flow: dense correspondence across scenes and its applications. IEEE Trans. Pattern Anal. Mach. Intell. **33**, 978–994 (2011)
32. Brox, T., Malik, J.: Large displacement optical flow: descriptor matching in variational motion estimation. IEEE Trans. Pattern Anal. Mach. Intell. **33**, 500–513 (2011)
33. Weinzaepfel, P., Revaud, J., Harchaoui, Z., Schmid, C.: DeepFlow: large displacement optical flow with deep matching. In: ICCV, pp. 1385–1392 (2013)
34. Hu, Y., Song, R., Li, Y.: Efficient coarse-to-fine patchmatch for large displacement optical flow. In: Proceedings of CVPR, pp. 5704–5712 (2016)
35. Lempitsky, V., Rother, C., Roth, S., Blake, A.: Fusion moves for markov random field optimization. IEEE Trans. Pattern Anal. Mach. Intell. **32**, 1392–1405 (2010)
36. Wedel, A., Pock, T., Zach, C., Bischof, H., Cremers, D.: An improved algorithm for TV-L1 optical flow. In: Proceedings of Statistical and Geometrical Approaches to Visual Motion Analysis, pp. 23–45 (2009)
37. Dosovitskiy, A., et al.: FlowNet: learning optical flow with convolutional networks. In: Proceedings of CVPR (2015)

38. Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T.: FlowNet 2.0: evolution of optical flow estimation with deep networks. In: Proceedings of CVPR (2017)
39. Sun, D., Yang, X., Liu, M.-Y., Kautz, J.: PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In: Proceedings of CVPR (2018)
40. Ranjan, A., Black, M.J.: Optical flow estimation using a spatial pyramid network. In: Proceedings of CVPR (2017)
41. Yu, J.J., Harley, A.W., Derpanis, K.G.: Back to basics: unsupervised learning of optical flow via brightness constancy and motion smoothness. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9915, pp. 3–10. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-49409-8_1
42. Liu, P., Lyu, M., King, I., Xu, J.: SelFlow: self-supervised learning of optical flow. In: Proceedings of CVPR (2019)
43. Wang, Y., Yang, Y., Yang, Z., Zhao, L., Wang, P., Xu, W.: Occlusion aware unsupervised learning of optical flow. In: Proceedings of CVPR (2018)
44. Hur, J., Roth, S.: Iterative residual refinement for joint optical flow and occlusion estimation. In: Proceedings of CVPR (2019)
45. Liu, P., King, I., Lyu, M.R., Xu, J.: DDFlow: learning optical flow with unlabeled data distillation. In: Proceedings of AAAI Conference on Artificial Intelligence, vol. 33, pp. 8770–8777 (2019)
46. Ashburner, J.: A fast diffeomorphic image registration algorithm. Neuroimage **38**, 95–113 (2007)
47. Blake, A., Zisserman, A.: Visual Reconstruction. MIT Press, Cambridge (1987)
48. Pratt, W.K., Kane, J., Andrews, H.C.: Hadamard transform image coding. Proc. IEEE. **57**, 58–68 (1969)
49. Lee, C.-Y., Xie, S., Gallagher, P., Zhang, Z., Tu, Z.: Deeply-supervised nets. In: Proceedings of MLR, pp. 562–570 (2015)
50. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: Proceedings of ICLR (2017)
51. Fischl, B.: FreeSurfer. NeuroImage. **62**, 774–781 (2012)
52. Hoffmann, M., Billot, B., Greve, D.N., Iglesias, J.E., Fischl, B., Dalca, A.V.: SynthMorph: learning contrast-invariant registration without acquired images. IEEE Trans. Med. Imaging **41**, 543–558 (2021)

Optimisation

Learn to Fuse Input Features for Large-Deformation Registration with Differentiable Convex-Discrete Optimisation

Hanna Siebert(✉) and Mattias P. Heinrich

Institute of Medical Informatics, Universität zu Lübeck, Lübeck, Germany
{siebert,heinrich}@imi.uni-luebeck.de

Abstract. Hybrid methods that combine learning-based features with conventional optimisation have become popular for medical image registration. The ConvexAdam algorithm that ranked first in the comprehensive Learn2Reg registration challenges completely decouples semantic and/or hand-crafted feature extraction from the estimation of the transformation due to the difficulty of differentiating the discrete optimisation step. In this work, we propose a simple extension that enables backpropagation through discrete optimisation and learns to fuse the semantic and hand-crafted features in a supervised setting. We demonstrate state-of-the-art performance on abdominal CT registration.

Keywords: Large deformation registration · Convex optimisation · End-to-end learning

1 Introduction and Related Work

While end-to-end learning of fully-convolutional networks is the method of choice for semantic segmentation, image registration continues to benefit from integrating conventional optimisation steps, e.g. pairwise instance optimisation [7], a discretised search of displacements [1] or iterative recurrent updates [9]. Discrete optimisation has been shown to yield excellent registration quality for numerous tasks [2,5,7] but does rely on non-differentiable steps which would prevent its use in end-to-end learning. We aim for a method that offers the possibility to use discrete optimisation in an end-to-end learning setting. Therefore, we introduce a differentiable convex discrete optimisation approach that is able to align images with large deformations. This differentiable optimisation is used to learn the fusion of semantic and hand-crafted image features.

2 Method

Figure 1 gives an overview of our method: First, hand-crafted and semantic features are extracted from the input images, concatenated and passed to a small

A. Hering et al. (Eds.): WBIR 2022, LNCS 13386, pp. 119–123, 2022.
https://doi.org/10.1007/978-3-031-11203-4_13

network comprising layers for feature fusion. The fixed and moving features output from this network are then used for our differentiable discretised convex optimisation method to align images with large deformations.

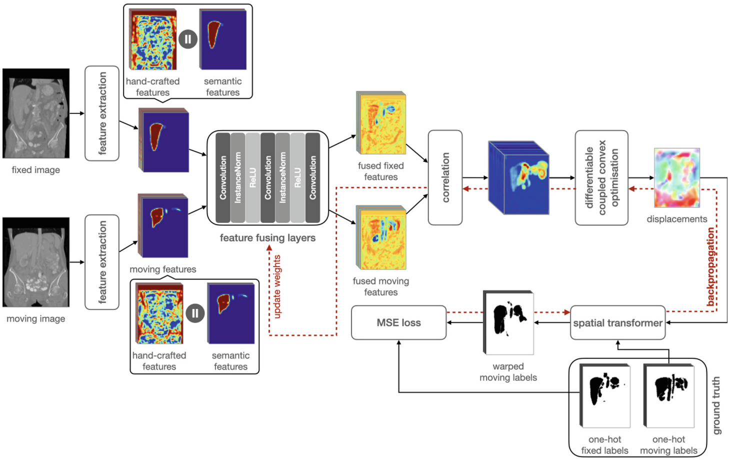


Fig. 1. Overview of our method: Hand-crafted and semantic features are concatenated and fused with feature fusing network layers. The fused features are used for our differentiable optimisation method to compute displacements. For backpropagation, warped moving and fixed labels are passed to a MSE loss function to update the feature fusing network's weights whereas the feature extraction part of the framework remains frozen.

2.1 Differentiable Convex-Discrete Optimisation

For pairwise deformable image registration, a deformation field $\mathbf{u}$ is sought that minimises the cost function $E(I_F, I_M, \mathbf{u})$ to align a fixed image I_F and a moving image I_M. In [3], a non-differentiable convex-discrete method has been proposed to find a deformation field $\mathbf{u}$ by solving a combined cost function

$$E(\mathbf{v}, \mathbf{u}) = DSV(\mathbf{v}) + \frac{1}{2\theta}(\mathbf{v} - \mathbf{u})^2 + \alpha|\nabla \mathbf{u}|^2 \tag{1}$$

that ensures similarity and smoothness optimisation. In this function, $\mathbf{v}$ is an auxiliary second deformation field used to compute the displacement space volume DSV. The regularisation parameter α controls the smoothness of the deformation field and the parameter θ models the coupling between similarity and regularisation penalty and is decreased during iterative solving of the equation. The optimal selection of $\mathbf{v}$ with respect to the similarity term can be performed globally optimal using local cost aggregation [3].

In this work, we introduce a differentiable discretised convex optimisation by replacing argmin operators with their corresponding softmin counterparts and make suitable adjustments to hyper-parameters that reduce memory requirements for end-to-end learning. Coupled-convex discrete optimisation [3] approximates more complex MRF-solutions by the following steps:

(0) initialisation of the current displacement field to zeros
(1) computation of a correlation volume based on sum of squared differences of feature tensors (the volume comprises 6 dimensions, 3 spatial dimension and 3 displacement dimensions)
(2a) a regularising coupling term that adds 3D parabolas in displacement dimensions that are rooted at the current displacement solution
(2b) the argmin operator (across all possible displacements) that defines a new regularised displacement field
(2c) a spatial smoothing step (e.g. a box-filter)

The correlation volume (step (1)) directly depends on the feature maps obtained from fixed and moving scans. By defining a large enough capture range and correspondingly a discrete mesh grid of relative displacements the method can robustly find a near global optimum without multiple warping steps or cascaded architectures. Steps (2a)–(2c) are iteratively repeated with a continuously increasing weight for the coupling term that helps to ensure convergence of the optimisation. Step (2b), which takes the argmin is not differentiable and will be replaced with a softmin operator along the displacement dimension followed by a point-wise multiplication with the relative displacements of the predefined discrete mesh grid and subsequent reduction.

2.2 Learning of Input Feature Fusion

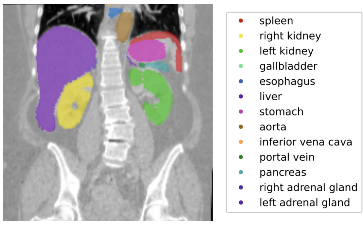
Previous work [3,7] has shown that hand-crafted MIND features [4] or automatic nnU-Net segmentations [6] can be used as input for a coupled convex optimisation method for image registration. In this work, we combine hand-crafted and semantic features by fusing them with help of trainable feature fusing network layers comprising two $1 \times 1 \times 1$-convolutions followed by instance normalisations and ReLU activations. The first convolution increases the number of feature channels to 32 and a third $1 \times 1 \times 1$-convolution reduces the number of feature channels to 15. The resulting feature maps are then used to solve the differentiable convex-discrete optimisation problem described in Sect. 2.1 in order to compute the displacement fields that are then used to warp the moving label maps. One-hot representations of warped and fixed label maps weighted inversely proportional to the square root of the class frequency are passed to a MSE loss function that is used to train the feature fusing network's parameters whereas the feature extraction part of the framework stays frozen.

3 Experiments and Results

For our experiments we use the Learn2Reg-2020 challenge's (task 3) dataset containing 30 abdominal inter-patient CT scans with 13 manually labeled abdominal

Table 1. Left: Quantitative results: Accuracy is measured by the Dice similarity of segmentations and the 95% Hausdorff distance for segmentations. Plausibility of the deformations is measured by the standard deviation of the logarithmic Jacobian determinant. Right: example visualisation of fixed image and warped moving labels.

	Dice [%]	HD [mm]	SDlogJ
initial	25.14	40.21	–
MIND features	37.79	37.22	0.050
nnU-Net features	50.56	24.71	0.021
concatenated features	49.71	28.33	0.050
fused features	56.37	24.13	0.049



organs and a resolution of 192 × 160 × 256 [5,10]. The scans have been linearly pre-registered and split into 20 training cases and 10 test cases. For evaluation we consider all possible pairwise combinations of the test cases. From the image data, we extract MIND features (leading to 12 feature channels) and compute one-hot encoded label features by applying a nnU-Net trained on the 20 training cases (leading to 14 feature channels). We downsample the features to a resolution of 48 × 40 × 64, concatenate them and pass the 26-channel input to our feature fusing network. The network's 15-channel output is then used for displacement computation with the differentiable convex optimisation method. Therefore, we use a displacement range that covers $\sim$ 32 mm within the scanned abdominal region and scale the softmin operation's output (step (2b)) by half of the downsampled feature dimensions. The feature fusing network is trained for 50 epochs using Adam and a learning rate of 0.005.

For evaluation, we upsample the obtained displacement fields to the original image resolution. We compare our fused features with the direct use of MIND features, nnU-Net label features, and concatenation of MIND and nnU-Net features. The results given in Table 1 show that the fusion of MIND and nnU-Net features clearly outperforms the other investigated feature variants with an average Dice score of 56.37% compared to 50.56% when using only nnU-Net features. As using nnU-Net features yields to a deformation field that is optimised to warp the foreground structures, the SDlogJ value is lower than when MIND features are involved. We evaluated the potential problem of label bias with an experiment on additional structures (lumbar and thoracic vertebrae[1] [8]) unseen for the nnU-Net segmentation training and our fusion learning. While using only MIND features yields the highest accuracy we see great potential for the proposed feature fusion that only reduced the Dice score of the spine by 5% while the nnU-Net-based registration results in a drop of 42%. Hence the influence of label bias is substantially reduced.

[1] https://github.com/MIRACLE-Center/CTSpine1K.

4 Discussion and Conclusion

This work introduced a differentiable version of coupled convex discrete optimisation for image registration with large deformation. It has opened up the possibility of end-to-end feature learning and has well-performed for our feature fusing network. We show that the fusion of semantic label features and hand-crafted features based on image self-similarities leads to an improved registration performance compared to either using only semantic or only hand-crafted features or the simple concatenation of both.

References

1. Dosovitskiy, A., et al.: FlowNet: learning optical flow with convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2758–2766 (2015)
2. Heinrich, M.P., Jenkinson, M., Brady, M., Schnabel, J.A.: MRF-based deformable registration and ventilation estimation of lung CT. IEEE Trans. Med. Imag. **32**(7), 1239–1248 (2013)
3. Heinrich, M.P., Papież, B.W., Schnabel, J.A., Handels, H.: Non-parametric discrete registration with convex optimisation. In: Ourselin, S., Modat, M. (eds.) WBIR 2014. LNCS, vol. 8545, pp. 51–61. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-08554-8_6
4. Heinrich, M.P., Jenkinson, M., Papież, B.W., Brady, S.M., Schnabel, J.A.: Towards realtime multimodal fusion for image-guided interventions using self-similarities. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013. LNCS, vol. 8149, pp. 187–194. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40811-3_24
5. Hering, A., et al.: Learn2reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning. arXiv preprint arXiv:2112.04489 (2021)
6. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)
7. Siebert, H., Hansen, L., Heinrich, M.P.: Fast 3D registration with accurate optimisation and little learning for Learn2Reg 2021. In: Aubreville, M., Zimmerer, D., Heinrich, M. (eds.) Biomedical Image Registration, Domain Generalisation and Out-of-Distribution Analysis. MICCAI 2021. LNCS, vol. 13166, pp. 174–178. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-97281-3_25
8. Smith, K., et al.: Data from CT_colonography. Cancer Imag. Arch. (2015). https://doi.org/10.7937/K9/TCIA.2015.NWTESAY1
9. Teed, Z., Deng, J.: RAFT: recurrent all-pairs field transforms for optical flow. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12347, pp. 402–419. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58536-5_24
10. Xu, Z., et al.: Evaluation of six registration methods for the human abdomen on clinically acquired CT. IEEE Trans. Biomed. Eng. **63**(8), 1563–1572 (2016)

Multi-magnification Networks for Deformable Image Registration on Histopathology Images

Oezdemir Cetin[1], Yiran Shu[1], Nadine Flinner[2], Paul Ziegler[2], Peter Wild[2], and Heinz Koeppl[1](✉)

[1] Department of Electrical Engineering and Information Technology, Technische Universität Darmstadt, Darmstadt, Germany
heinz.koeppl@tu-darmstadt.de

[2] Senckenberg Institute of Pathology, University Hospital Frankfurt, Frankfurt, Germany

Abstract. We present an end-to-end unsupervised deformable registration approach for high-resolution histopathology images with different stains. Our method comprises two sequential registration networks, where the local affine network can handle small deformations, and the non-rigid network is able to align texture details further. Both networks adopt the multi-magnification structure to improve registration accuracy. We train the proposed networks separately and evaluate them on the dataset provided by the University Hospital Frankfurt, which contains 41 multi-stained histopathology whole-slide images. By comparing with methods using the single-magnification structure, we confirm that the proposed multi-view architecture can significantly improve the performance of the local affine registration algorithm. Moreover, the proposed method achieves high registration accuracy of contents at the cell level and is potentially applicable to other medical image alignment tasks.

Keywords: Histopathological image · Affine transformation · Non-rigid registration · Unsupervised learning · Multi-magnification network

1 Introduction

Histopathological whole slide images, i.e., digital tissue slides produced by scanning conventional glass slides under high-resolution microscopy, are vital for modern histopathology analysis [15]. Standard whole slide images employ the pyramid structure to support different resolutions, making it easy for pathologists to observe by zooming. Each layer of the pyramid corresponds to a resolution level, with the bottom being the highest resolution information. In general, histopathologists utilise various staining techniques based on chemical features of the tissue, e.g. Hematoxylin-Eosin (H&E), periodic-acid Schiff (PAS) or elastic-van Gieson (EvG). In addition, antibody-mediated visualization of specific proteins, termed immunohistochemistry, is widely used in modern histopathology.

A. Hering et al. (Eds.): WBIR 2022, LNCS 13386, pp. 124–133, 2022.
https://doi.org/10.1007/978-3-031-11203-4_14

As tissue specimens are prepared by approx. 3 µm-thin cuts each specimen represents an unique sample and slides obtained from directly adjacent tissue differ slightly in their morphology. Even when the same tissue slide is used for multiple staining, e.g. by bleaching and re-staining, shifts and/or deformations inevitably occur. These digital multi-stained histopathology images that are not aligned accurately pose obstacles to the diagnosis or further processing, thus need to be registered first.

Image registration is the process of matching two images geometrically so that corresponding coordinate points in both images correspond to the same physical region of the scene being imaged [21]. Biomedical image registration constitutes one of the key research areas for medical analysis that has been extensively studied. Traditional registration methods search for spatial transformation that brings the defined similarity metric to be optimum by an iterative optimization algorithm [1]. Nevertheless, the superiority of accuracy and robustness of classical approaches come at the cost of time, which becomes the main bottleneck in archiving desirable performance for practical applications. With the revival of deep learning, attempts have been made to develop learning-based approaches to implement faster registration, which can be grouped into three main categories [5,6]: (*i*) deep iterative registration, which follows the framework of traditional methods but instead adopts similarity metrics learned by deep neural networks [16,17], (*ii*) supervised transformation prediction, utilizing the known ground truth transformations to define the cost function [9,18], (*iii*) Unsupervised transformation prediction, where a spatial transformation network is applied to calculate the error of the given metric(s) with an appropriate regularization term [2,19]. The first class of methods inherits the time-consuming drawback of conventional approaches due to the iterative process, whereas the supervised training requires a large amount of data with annotations. In contrast, unsupervised transformation approaches produce the supervisory signals required for training directly by data and can achieve real-time registration during prediction. Therefore, we focused on the unsupervised methods in this work.

An obstacle to applying learning-based methodologies to histopathology images concerns their ultra-high resolution. Some studies have resampled images down to an acceptable memory limit before deformation estimation [3,15]. However, such detailed information as the cell morphological structure is almost impossible to observe on low-resolution images, becoming a key hamper in improving alignment accuracy. An alternative solution is to perform registration on smaller patches [8,12]. The shortcoming of this approach is the irreversible loss of neighboring information when splitting the images, resulting in the narrow field-of-view. In this work, we propose two deep multi-magnification network architectures for patch-based affine and non-rigid registration. The proposed local affine algorithm can effectively deal with imperceptible collective shifts of cell nuclei in the low-resolution pattern, and non-rigid registration is able to align further the cell components that are slightly altered in the morphological structure. We train the presented networks unsupervised and yield higher registration accuracy than the methods using only ordinary single-magnification networks. The result reaches precise alignment at the cellular level under the maximum resolution of histopathology WSIs, which significantly contributes to the manual/automatic pathological diagnosis on the differently stained tissue sections.

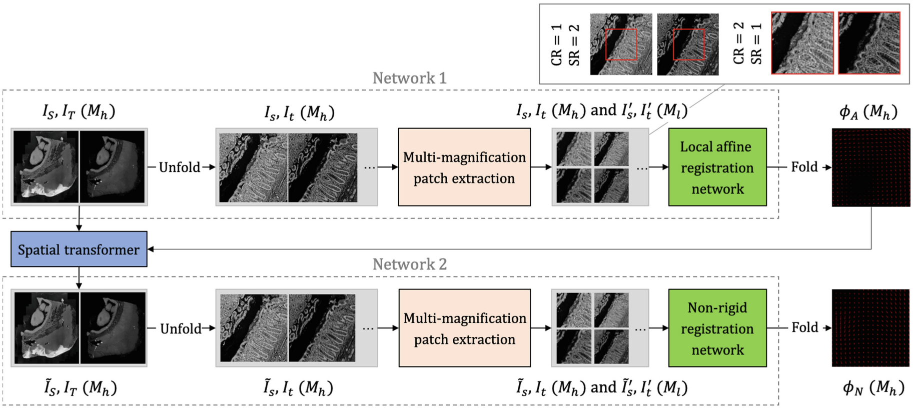

CR = 1
SR = 2
CR = 2
SR = 1
Network 1
$I_S, I_T\ (M_h)$
Unfold
$I_s, I_t\ (M_h)$
Multi-magnification patch extraction
$I_s, I_t\ (M_h)$ and $I_s', I_t'\ (M_l)$
Local affine registration network
Fold
$\phi_A\ (M_h)$
Spatial transformer
Network 2
Unfold
Multi-magnification patch extraction
Non-rigid registration network
Fold
$\tilde{I}_S, I_T\ (M_h)$
$\tilde{I}_s, I_t\ (M_h)$
$\tilde{I}_s, I_t\ (M_h)$ and $\tilde{I}_s', I_t'\ (M_l)$
$\phi_N\ (M_h)$

Fig. 1. Overview of the proposed algorithm: Both networks take as input concatenated patches I_s, I_t for M_h (high) magnification, and concatenated patches $I_s{}'$, $I_t{}'$ for M_l (low) magnification. An example in the upper right corner illustrates the construction process of a patch set, where the cropping rate (CR) and sampling factor (SR) used to build patches for each magnification level are given. The red boxes denote the corresponding regions at different magnifications. (Color figure online)

2 Methods

Let I_S, I_T: $\Omega \rightarrow \mathbb{R}$ represent the whole slide source and target images, defined in the spatial domain $\Omega \subset \mathbb{R}^d$, where d denotes ($d = 2$ in this study) spatial dimensionality of the given data. Similarly, I_s, I_t: $\omega \rightarrow \mathbb{R}$ with $\omega \subset \Omega$ represent the patch-wise source and target images, extracted from I_S and I_T. Assuming that the image pairs to be registered are pre-aligned well, we aim to find two deformation fields ϕ_A, ϕ_N: $\Omega \rightarrow \Omega$ to deform the source image such that:

$$I_S(\phi_N \circ \phi_A(x)) \approx I_T, \forall x \in \Omega. \tag{1}$$

Here "$\circ$" represents the composition of deformations and $I(\phi)$ indicates I deformed by ϕ. The deformations ϕ_A, ϕ_N are defined as a patch-wise affine deformation and a pixel-wise non-rigid deformation, respectively. They are obtained by aggregating the local deformations ϕ_a^p, ϕ_n^p : $\omega \rightarrow \omega$ of image patches (I_s, I_t) extracted from (I_S, I_T), where p indicates the index of the patch on the whole slide image. Two convolutional neural networks f_a and f_n are used to realize the affine registration $\phi_a = f_a(I_s, I_t)$ and non-rigid registration $\phi_n = f_n(I_s(\phi_a), I_t)$, respectively.

An affine registration network is leveraged to learn the affine transformation $\phi_a := Tx$, where $T \in \mathbb{R}^{d \times m}$ with $m = d + 1$. Next, the affinely registered images are fed into the non-rigid registration network to learn the displacement field $u(x)$ with $\phi_n := x + u(x)$, which represents the displacements for $\forall x \in \omega$ in the vertical and horizontal directions.

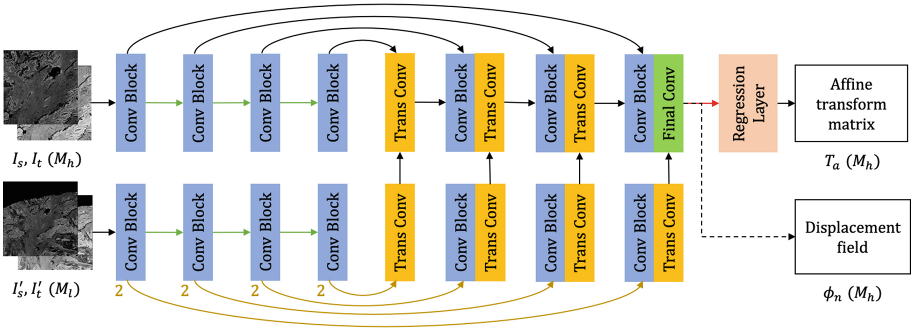


Fig. 2. Architecture of the local affine and non-rigid registration networks: *Conv Block* includes two sets, each consisting of a 3×3 convolution layer with group normalization (GN), activated by PReLU. *Trans Block* comprises a 2×2 transposed convolution layer with a stride of 2 followed by GN and PReLU activation. The green and red arrows indicate maximum pooling and average pooling, respectively. The center cropping operations are denoted by brown arrows with the cropping rates written in brown. Other blocks are described in the text. (Color figure online)

The input of both networks is a set of image patches with different magnifications, providing multiple field-of-views to the networks. Figure 1 offers an overview of the proposed registration algorithm for the case of two magnification levels. The strategy adopted for extracting multi-magnification patches in this work is described as follows: In a multi-magnification set, all other patches are obtained by center-cropping the base image with different cropping rates. Then, the patches are downsampled with the corresponding sampling factors to uniform the patch size. The downsampled base image is the one with the lowest magnification level in the set. Registration networks take the patch set as input and predict the local affine transform matrix/displacement field corresponding to the patch with the highest magnification level, as details described in the next section. According to Eq. 1, the final deformation for the given images I_S and I_T is obtained by composing the folded ϕ_A and ϕ_N.

2.1 Network Architectures

The proposed networks are inspired from [7], which contains multiple magnification layers that obtain more information from different field-of-views. Considering that the architectures of both networks are quite similar, they are shown in one figure for brevity, as visualized in Fig. 2. The concatenation of the high-magnification patches I_s and I_t is fed into the target magnification layer based on the U-Net [13], to extract the higher magnification feature maps. During reconstruction, these feature maps are concatenated with the corresponding lower magnification feature maps extracted from the lower-magnification patches ${I_s}'$ and ${I_t}'$ in another magnification layer. To limit the usage of feature maps from cropped boundary areas in a wide field-of-view, the lower magnification feature maps are center-cropped with a given cropping rate followed by up-sampling utilizing transpose convolution to match the size.

In the local affine network, $FinalBlock$ has the same structure as $ConvBlock$ but a stride of 2, followed by an adaptive average pooling layer. The reconstructed feature maps are transformed into six numeric parameters through a fully-connected layer and then rearranged into the resulting affine transform matrix T in the regression layer. Whereas, in the non-rigid network, the reconstructed feature maps are compressed utilizing $Final\ Block$, a stack of a 3×3 and a 1×1 convolution layer, into two-channel displacement field $u(x)$.

2.2 Loss Function

Assume that $\phi : \omega \rightarrow \omega$ is the local deformation field estimated by networks with image patches I_s and I_t as input, the loss function can be described as

$$\mathcal{L}\left(I_s, I_t, \phi\right) = \mathcal{L}_S\left(I_s(\phi), I_t\right) + \lambda \mathcal{L}_R\left(\phi\right), \tag{2}$$

where the first term $\mathcal{L}_S$ measures the similarity between the warped source and the target patches, and $\mathcal{L}_R$ is a regularization term considered only in the non-rigid network. Parameter λ controls the trade-off between these two terms as a hyperparameter in the training process.

We choose the normalized cross-correlation (NCC) [10] as the similarity metric $\mathcal{L}_S$. Let I_1, I_2 be two images then this similarity can be computed as

$$NCC(I_1, I_2) = \frac{1}{N-1} \sum_{x \in \omega} \frac{(I_1(x) - \bar{I_1})(I_2(x) - \bar{I_2})}{\sigma_{I_1} \sigma_{I_2}}, \tag{3}$$

where N indicates the number of non-zero pixels, $\bar{I}$ and σ_I represent the mean and standard deviation of the intensities in image I, respectively. The negative normalized cross-correlation (NCC) is used in training to minimize the loss function, while a higher NCC value corresponds to a higher similarity between images.

Under the intuition that a desirable deformation field should not vary too much between nearby points, the curvature regularization [4] is used to constrain the geometric smoothness of the displacement field ϕ predicted by the non-rigid network, i.e.,

$$\mathcal{L}_R(\phi) = \sum_{x \in \omega} \parallel \nabla \phi(x) \parallel^2 . \tag{4}$$

3 Experiments

The University Hospital Frankfurt (UKF) provided the images used in this study to evaluate the proposed algorithm, with clinical data removed and completely anonymized. The UKF dataset comprises two parts: The first part offers 36 histopathological WSIs, where every two images are from the same tissue section, respectively stained with H&E and IHC-CD8. The second part consists of 5 WSIs obtained from two staining experiments in which multiple staining was performed on the tissue slides from one tissue in different orders. All WSIs are provided as .mrxs files with a unified specification. Each of them contains images at nine resolutions with a downsampling factor of 2, where the full resolution exceeds $180k \times 90k$ pixels in size. We generated 18 and 5 image pairs respectively from two parts of the UKF dataset for training and evaluation. The experiment details are presented next.

3.1 Experimental Settings

Data Preprocessing. We removed large background areas in the raw data by a boundary detection algorithm and then converted them into single-channel grayscale images. The rigid alignment method derived from [20] was adopted to handle the large misalignment of the image pairs.

Technical Details. The proposed algorithm was implemented by modifying and extending the DeepHistReg framework [20]. Unsupervised methods were trained on the resolution-level 4 images whose size varies from $3k$ to $7k$ pixels in one dimension. The images are split into overlapping patches, followed by extracting 224×224 patches of different magnification levels as the input to the networks. We trained both presented networks with a batch size of 4 using Nvidia Tesla P100 (PCIe). The Adam optimizer with an initial learning rate of $1e{-}3$ and a decay rate of 0.95 was adopted to update the network parameters. The constraint coefficient λ for the non-rigid network training was chosen to be 60.

Baseline Methods. We built two single-magnification networks for local affine and non-rigid registration as the baseline models for comparison. The architecture of both networks inherited the target magnification layer of the corresponding multi-amplification network with some adaptations. The training settings were the same as the proposed methods.

3.2 Evaluation Metrics

We quantified the registration accuracy by several similarity metrics since no ground truth such as landmarks or segmentation maps are provided for the UKF dataset. Except for the metric NCC used as the objective function during network training, the quality of the deformation fields was also evaluated by the

Mean-Squared-Error (MSE) [11] and the normalized Mutual-Information (NMI) [14], which are respectively defined as

$$MSE(I_1, I_2) = \frac{1}{N} \sum_{x \in \omega} \left[I_1(x) - I_2(x)\right]^2, \tag{5}$$

$$NMI(I_1, I_2) = \frac{2 \cdot H(I_1, I_2)}{H(I_1) + H(I_2)}, \tag{6}$$

where H indicates Shannon's entropy and $H(I_1, I_2)$ represents the dependence of variables (images) I_1 and I_2.

Table 1. Comparison among methods with single/multi-magnification registration networks, containing the average inference time and performance quantified by the similarity metrics NCC, MSE, and NMI (arrows indicate the trend of the increased similarity): The methods are named according to the adopted network architectures, where S/M stands for networks with the single/multi-magnification structure, and A/N denotes the local affine transformation and non-rigid deformation. For example, MASN refers to combining a multi-magnification local affine network and a single-magnification non-rigid network. Besides, an iterative approach is applied based on the presented method, with the number of iterations denoted in parentheses.

Metric	Initial	SASN	SAMN	MASN	MAMN	MASN(3)
NCC ↑	0.6828	0.7123	0.7060	**0.7461**	0.7443	**0.7728**
MSE ↓	0.0403	0.0376	0.0382	**0.0336**	0.0338	**0.0305**
NMI ↑	0.1670	0.1781	0.1756	0.1952	**0.1954**	**0.2038**
Time (sec)	–	25.48	28.96	28.95	32.31	45.29

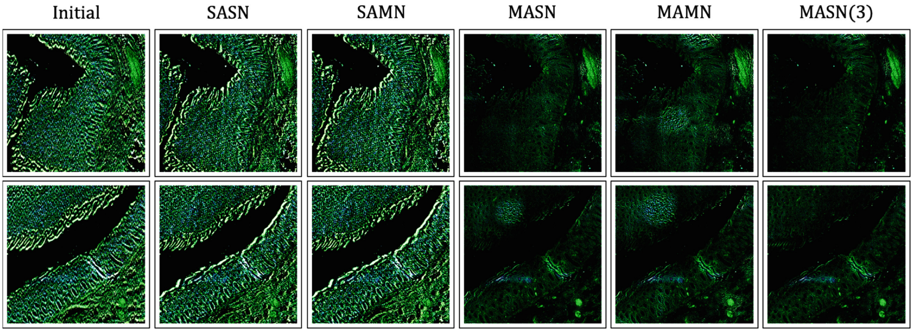


Fig. 3. Local subtractions of a high-resolution image pair registered by different methods: The non-overlapping regions appear as fluorescent green due to the nature of stains. For visibility, the contrast/brightness of images has been increased by 50%. (Color figure online)

4 Results

Table 1 summarizes the overall performance of our proposed algorithm in comparison to approaches containing one or more baseline models. All of them take images pre-aligned by rigid alignment as input.

As shown in Table 1, the proposed multi-magnification structures outperformed the ordinary single-magnification architecture for the local affine algorithm with remarkable benefits, whereas yielding almost no improvement in the performance of the non-rigid network. The increase in runtime due to the multi-magnification structure is not significant compared to the base runtime (SASN). According to the proposed algorithm, the difference in time will decrease exponentially for smaller image pairs. By iterating the prediction on the previous result by the same network, we obtained registration results with significantly higher accuracy.

We upsampled the predicted deformation fields for generating the registered images at a higher resolution. By performing local subtraction between the deformed source and target images, we evaluated the registration performance of different methods at the cellular level, as illustrated in Fig. 3. It can be observed that the local affine network improved by the multi-magnification structure is crucial for the enhancement of the overall performance. The cell nuclei can overlap completely in the best cases.

5 Discussion and Conclusion

In modern histopathology multiple staining techniques are used to detect specific structures within biological tissues. Each technique highlights different characteristics of the tissue and proper analysis needs to address the spatial distribution of these characteristics. In this context, we developed two novel deep networks with the multi-magnification structure for patch-based image registration, which can learn peripheral information outside the patches as auxiliary information to improve network performance. The presented method is of great importance for biomedical image registration since studies for them can often be performed only on smaller patches due to the large image size. Moreover, the network architectures can be easily expanded with more magnification levels. Nevertheless, this expansion makes little sense since too many field-of-views may instead negatively affect the network performance, especially for cases with no apparent global misalignment.

Our experiments compared the impact of single- and multi-magnification networks on the overall alignment performance by different network combinations. The results revealed that the multi-magnification structure could significantly improve the performance of the patch-based affine registration network. However, it yielded little success on the local non-rigid network. This might mainly attribute to the transformation nature of these two registration methods. The lack of neighboring information can aggravate the estimation error of deformation for the whole image patch region by the local affine approach, while this error occurs only within the edge region of the image patches in the non-rigid method due to the dense prediction. Therefore, the enhancement of the non-rigid method by the multi-magnification structure was much less evident than that of the local affine approach. Besides, we introduced an iterative approach on the method with the best performance, which further improved the registration accuracy, with an acceptable growth of inference time. The proposed method has the potential to be applicable for other medical image registration tasks.

Acknowledgement. H.K. acknowledges support from the European Research Council (ERC) with the consolidator grant CONSYN (nr. 773196). O.C. is supported by the Alexander von Humboldt Foundation Philipp Schwartz Initiative.

References

1. Costin, H.N., Rotariu, C.: Registration of multimodal medical images. Comput. Sci. J. Moldova **51**(3), 231–254 (2009)
2. Dalca, A.V., Balakrishnan, G., Guttag, J., Sabuncu, M.R.: Unsupervised learning for fast probabilistic diffeomorphic registration. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11070, pp. 729–738. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00928-1_82
3. Feuerstein, M., Heibel, H., Gardiazabal, J., Navab, N., Groher, M.: Reconstruction of 3-D histology images by simultaneous deformable registration. In: Fichtinger, G., Martel, A., Peters, T. (eds.) MICCAI 2011. LNCS, vol. 6892, pp. 582–589. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-23629-7_71
4. Fischer, B., Modersitzki, J.: A unified approach to fast image registration and a new curvature based registration technique. Linear Algebra Appl. **380**, 107–124 (2004)
5. Fu, Y., Lei, Y., Wang, T., Curran, W.J., Liu, T., Yang, X.: Deep learning in medical image registration: a review. Phys. Med. Biol. **65**(20), 20TR01 (2020)
6. Haskins, G., Kruger, U., Yan, P.: Deep learning in medical image registration: a survey. Mach. Vis. Appl., 1–18 (2020). https://doi.org/10.1007/s00138-020-01060-x
7. Ho, D.J., et al.: Deep multi-magnification networks for multi-class breast cancer image segmentation. Comput. Med. Imaging Graph. **88**, 101866 (2021)
8. Lotz, J., et al.: Patch-based nonlinear image registration for gigapixel whole slide images. IEEE Trans. Biomed. Eng. **63**(9), 1812–1819 (2015)
9. Lv, J., Yang, M., Zhang, J., Wang, X.: Respiratory motion correction for free-breathing 3d abdominal MRI using CNN-based image registration: a feasibility study. Br. J. Radiol. **91**(xxxx), 20170788 (2018)
10. Modersitzki, J.: FAIR: flexible algorithms for image registration. SIAM (2009)
11. Pishro-Nik, H.: Introduction to Probability, Statistics, and Random Processes. Kappa Research, Athens (2016)
12. Pitiot, A., Bardinet, E., Thompson, P.M., Malandain, G.: Piecewise affine registration of biological images for volume reconstruction. Med. Image Anal. **10**(3), 465–483 (2006)
13. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28

14. Schütze, H., Manning, C.D., Raghavan, P.: Introduction to Information Retrieval, vol. 39. Cambridge University Press, Cambridge (2008)
15. Schwier, M., Böhler, T., Hahn, H.K., Dahmen, U., Dirsch, O.: Registration of histological whole slide images guided by vessel structures. J. Pathol. Inform. **4**(Suppl) (2013)
16. Simonovsky, M., Gutiérrez-Becker, B., Mateus, D., Navab, N., Komodakis, N.: A deep metric for multimodal registration. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9902, pp. 10–18. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46726-9_2
17. So, R.W., Chung, A.C.: A novel learning-based dissimilarity metric for rigid and non-rigid medical image registration by using Bhattacharyya distances. Pattern Recogn. **62**, 161–174 (2017)
18. Uzunova, H., Wilms, M., Handels, H., Ehrhardt, J.: Training CNNs for image registration from few samples with model-based data augmentation. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10433, pp. 223–231. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66182-7_26
19. Wodzinski, M., Müller, H.: Unsupervised learning-based nonrigid registration of high resolution histology images. In: Liu, M., Yan, P., Lian, C., Cao, X. (eds.) MLMI 2020. LNCS, vol. 12436, pp. 484–493. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59861-7_49
20. Wodzinski, M., Müller, H.: DeephistReg: unsupervised deep learning registration framework for differently stained histology samples. Comput. Methods Programs Biomed. **198**, 105799 (2021)
21. Zitova, B., Flusser, J.: Image registration methods: a survey. Image Vis. Comput. **21**(11), 977–1000 (2003)

Realtime Optical Flow Estimation on Vein and Artery Ultrasound Sequences Based on Knowledge-Distillation

Till Nicke[1,4](✉), Laura Graf[2], Mikko Lauri[1], Sven Mischkewitz[3], Simone Frintrop[1], and Mattias P. Heinrich[2]

[1] Department of Informatics, University of Hamburg, Hamburg, Germany
{till.nicke,mikko.lauri,simone.frintrop}@uni-hamburg.de
[2] Institute of Medical Informatics, University of Lübeck, Lübeck, Germany
{graf,heinrich}@imi.uni-luebeck.de
[3] ThinkSono GmbH, Potsdam, Germany
sven@thinksono.com
[4] Fraunhofer Institute for Image Computing MEVIS, Lübeck, Germany

Abstract. In this paper, we propose an approach for realtime optical flow estimation in ultrasound sequences of vein and arteries based on knowledge distillation. Knowledge distillation is a technique to train a faster, smaller model by learning from cues of larger models. Mobile devices with limited resources could be key in providing effective point-of-care healthcare and motivate the search of more lightweight solutions in the deep learning based image analysis. For ultrasound video analysis, motion correspondences of image contents (anatomies) have to be computed for temporal context and for real time application, fast solutions are required. We use a PWC-Net's [1] optical flow estimation output to create soft targets to train a PDD-Net [2] as lightweight optical flow estimator. We analyse the students' performance on the challenging task of fast segmentation propagation of vein and arteries in ultrasound images. Experiments show that even though we did not fine-tune the teachers on this task, a model trained with soft targets outperformed a model trained directly with labels and without a teacher.

Keywords: Knowledge distillation · Realtime video inference · Ultrasound images

1 Introduction

The analysis of objects in a sequence of images is a task that plenty of research has been done for, recently mostly in the deep learning field [3]. To achieve a coherent and accurate result over the different time points, it is important that

A. Hering et al. (Eds.): WBIR 2022, LNCS 13386, pp. 134–143, 2022.
https://doi.org/10.1007/978-3-031-11203-4_15

the analysis of the current image considers the past. One way to represent this temporal context is in the form of estimated optical flow. However, the classical methodology for its' calculation is an iterative approach [4] too slow for realtime inference. Most recent image registration approaches based on deep learning (e.g. [5]) are computationally too expensive to be executed on mobile devices in the required time. Realtime estimation of optical flow of ultrasound sequences would be advantageous in many practical point-of-care ultrasound (POCUS) applications that are based on intelligent guidance through image analysis. The aim of this work is to train a network, that learns from larger, pre-trained flow estimation networks and is able to accurately propagate relevant information (e.g. segmentations of important anatomies) in ultrasound. Ultrasound images often exhibit ambiguous structure depiction and a network, that employs only 2D convolution without temporal context, is not able to perfectly interpret the image with satisfying accuracy. So instead, utilising the motion of the images can leverage temporal context without requiring access to the whole temporal sequence. A CNN can be trained to estimate the temporal context e.g. by learning to propagate anatomical labels correctly between two images (which is usually coined weakly-supervised registration [6]). Clinically, this is relevant e.g. for the diagnosis of deep vein thrombosis (DVT), for which vessels in the leg need to be labelled.

2 Related Work

2.1 Dynamic Ultrasound Analysis

The use of automated image analysis for ultrasound is constantly increasing both in research and practical clinical translations [7]. The recent MICCAI challenge CLUST [8] has studied the quality of image registration algorithms for tracking ultrasound but without realtime constraints. A Siamese network for respiratory motion estimation on ultrasound images has been proposed by Liu and colleagues [9], which is capable of tracking landmarks through a video sequence.

A system for compression-based DVT examination in ultrasound (US) images was proposed by Tanno and colleagues [10]. The system, named AutoDVT, uses a dual-task network to help make predictions about the patient's VTE status. One of the tasks consists of classifying the compression status of a registered vein as either closed or open. The network itself uses stacked consecutive frames as input to create temporal consistency. The different task networks share the majority of convolutional layers and only separate the two tasks in the last convolutional layer, thus each task regulates the other during training.

To achieve higher temporal consistency and capture a more holistic view of dynamic sequences, optical flow estimation between frames can be leveraged. To ensure fast inference time, it is of importance that the optical flow prediction takes as little time as possible, while still generating accurate estimations.

2.2 Optical Flow Estimation

In recent research in deep learning and optical flow estimation numerous capable network solutions have been proposed, including Flownet [11], its evolution Flownet2 [12], and PWC-Net [1]. Flownet uses CNN feature extractors on two images, correlates these features over a discretised displacement search window (originally 21 × 21 pixels with a stride of 4), and further processes these correlations to predict a flow field. Flownet2 extends the original Flownet approach by employing multiple different and fine-tuned versions of this architecture.

PWC-Net, which was proposed by Sun et al. [1], on the other hand, uses pyramidal images with a combination of a cost-volume layer and a warping layer to estimate the optical flow of the input images.

In the medical domain LapIRN [13] and PDD-Net [2] are two capable networks for estimating large deformations. PDD-Net utilizes deformable convolution layers for feature extraction, which are then correlated. The correlation layer is followed by a min convolution and mean-field inference to predict dense displacement probabilities in volumes.

Some of these networks are larger, with up to 162 million parameters and up to 0.6 s of inference time on an NVIDIA graphics card [11,12]. However, these models are very accurate, which makes them valuable teachers in a student-teacher setting. Other models, such as the PDD-Net, with less parameter counts use little space and computation.

2.3 Knowledge Distillation

Student-teacher learning, also known as knowledge distillation (KD), was proposed by Li et al. [14]. The method uses one (or more) large and accurately trained neural network(s), also called teacher, and tries to teach the output distribution to a smaller network, also called student, by minimizing the KL divergence between the teacher's output and the students' prediction.

Yuan et al. proposed that not only accurate teachers can be used in a knowledge distillation setting. In [15] they found that also insufficiently trained teachers can increase the performance of the students, as they provide a representative distribution of the classes in the classification task. Thus, the teachers not only provide accurate information about the output but also provide regularized soft targets.

In [16] Kim et al. compared the KL divergence as a loss function, which is widely used in knowledge distillation, to a mean squared error loss and found, that the mean squared error loss is superior to the KL divergence, especially, when using a small tau, as the label noise is mitigated.

2.4 Contributions

We utilize the aforementioned knowledge distillation process [14] to train a small and lightweight optical flow estimator network (PDD-Net) for ultrasound motion estimation and vessel segmentation propagation in ultrasound images. We also compare this method to a label loss trained network to evaluate the usage of the distilled knowledge and find an increase in Dice score, as well as a decrease in Hausdorff distance (HD). As segmented medical reference data is scarce, this approach could potentially help increase performances for ultrasound image processing.

We aim at a short inference time of the optical flow to either create an additional input for further image analysis networks or to use the optical flow itself for segmentation propagation on mobile devices, such as tablets or phones. This constrains size and throughput of the network, as computational power on mobile devices differs greatly from stationary setups. Therefore, we use a lightweight version of the aforementioned PDD-Net as student.

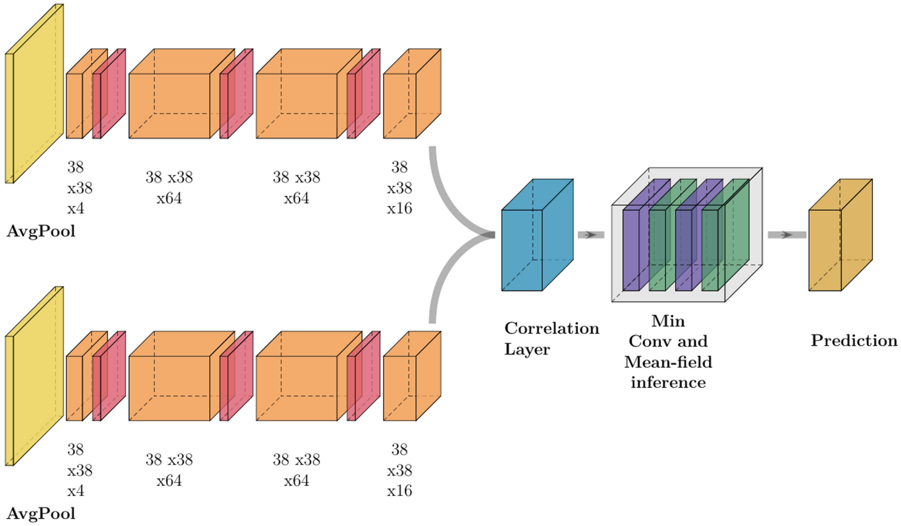


Fig. 1. Overview of the PDD-Network architecture for image registration, which comprises deformable convolutions with batch normalisation and ReLU (red), a correlation layer (blue) and differentiable mean-field inference as regularisation (purple and green). (Color figure online)

We use the PDD-Net [2], which achieved competitive results in the Learn2Reg challenge [17], and was made available[1] in a 2D version (Fig. 1). In this version of the model, an average pooled (yellow) input image is processed by three convolutional layers each followed by batch normalisation and ReLU (red). After the first convolution, we adapt the 2D implementation by applying an Obelisk layer [18], which is then followed by two more convolutional layers. The Obelisk layer is a form of deformable convolution, which uses learnable weights and a gridsampling operator, to increase the receptive field of the next convolutional layer [18].

[1] https://www.kaggle.com/mattiaspaul/learn2reg-tutorial.

For a fixed and a moving image, the extracted features are correlated, akin to the correlation used in Flownet-C [11] and then further processed with min convolutions and mean-field inference [19] (gray box).

The whole model yields an inference time of around 2.7 ms on an Nvidia RTX 2060 Ti GPU. When looking at the model (Fig. 1), we can see two feature extractors, which share weights. By processing one fixed frame at time t and keeping this frame as a fixed frame, we only need to process the moving frame at point $t + x$ of the video through the CNN. By reducing the convolutional operations needed during video processing, the network's inference time can be reduced to 1.7 ms. The same optimization can be applied when using different fixed images. In that case the extracted feature map of the moving frame (at time t) can be re-purposed as feature map of fixed frame (at time $t + x$), when a new moving frame is presented.

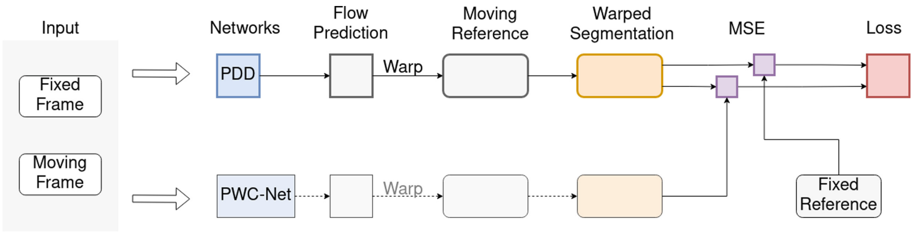


Fig. 2. Illustration of our concept for knowledge distillation for DL-based optical flow estimation. The teacher (PWC-Net) was not trained on ultrasound sequences but can provide a soft target for our student (PDD-Net) based on only a single reference frame segmentation.

The PDD-Net is trained on a combination of soft and hard targets. The hard target loss is calculated as the MSE between the one-hot encoded reference segmentation ("fixed reference" in Fig. 2), and the networks' prediction. The prediction is generated by using the predicted flow field to warp the reference segmentation from the moving frame towards the fixed frame. This warped segmentation is then compared to the reference segmentation of the fixed frame.

We use the established optical flow estimator PWC-Net [1] as a teacher to provide soft targets during training. This is done as shown in Fig. 2. To generate the soft targets, the PWC-Net's optical flow prediction is used to warp the reference segmentation of the moving frame towards the fixed frame. We calculate the MSE loss between the one-hot encoded warped moving reference segmentations of teacher and student networks. The soft and hard target loss are then summed up, where the soft target loss is scaled by 0.5.

Experimental Setup: We train two networks with different methods on the same data. One network is trained solely on hard labels (labeled PDD), as described above. The other network is trained with additional soft target influence (labeled PDD_{KD}). The dataset used for training and evaluation was pro-

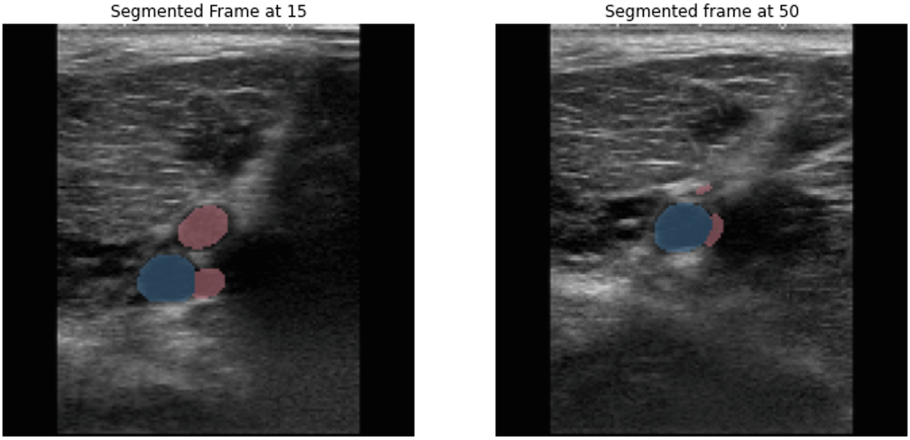


Fig. 3. Exemplary image pair used in the fine tuning data set. Reference segmentations are added for better visualization where the artery is shown in blue and the vein is marked in red. (Color figure online)

vided by ThinkSono GmbH[2]. It contains video sequences of DVT examinations that were annotated by experts. An overlay of these reference segmentations can be seen in Fig. 3. We use 250 video IDs to create two datasets with which we capture two distinct properties. The first dataset is created as a training dataset and contains 1743 image pairs with a fixed frame distance of 6 frames that were randomly sampled. Thus, capturing smaller and larger displacements while also providing heterogenous image quality. The second dataset is created as a fine-tuning dataset. This dataset is created to provide task specific data. For every ID, we select one random frame in the first fifth of the video, or before the onset of the vein compression (whatever came first). We then sample the coming frames with a frame distance of 4 and create various image pairs with the same fixed frame and different moving frames, resulting in 3285 image pairs. In this dataset larger displacements and vein compressions are captured. The evaluation task is to propagate a single reference segmentation of veins and arteries through a video of unseen IDs of about 10 s of a DVT examination.

We proceed to train the PDD-Net adaptation on the training data set with additional soft targets from the PWC-Net (Fig. 2) over 100 epochs with a learning rate of 0.002 and an Adam optimizer. We then trained the distilled network on the fine-tuning data set for 200 epochs with a learning rate of 0.00025. For comparison, we also train one version of the PDD-Net adaptation without additional soft targets in the same manner.

[2] https://thinksono.com/.

3 Results and Discussion

We evaluate both networks on 23 unseen videos containing approximately 1600 Frames overall. For each video, we selected one random frame in the first fifth of the video, which we refer to as f_t, for frame at time point t. Each following frame f_{t+x} is used as moving frame input. The estimated optic flow between f_t and f_{t+x} is used to warp the reference segmentation from t to $t+x$, where it is compared to reference segmentation at time pint $t+x$.

This procedure allows us to apply the mentioned runtime optimization towards video processing. By passing the fixed frame once, keeping it in memory for correlation, solely the moving frames need to be passed through the CNN for feature extraction. The reduced inference time per image is about as fast as a reference segmentation network, nnU-Net, which takes 1.6 ms on the same GPU (Nvidia RTX 280Ti).

As mentioned by Reinke [20] there are common limitations when applying only one metric to measure the performance of segmentation masks. Therefore, we evaluate the two networks on Dice score and Hausdorff distance. The dice score is used as a measurement of overlap between the reference and predicted segmentation. It ranges from 0 to 1, where 1 is the best score, which we have denoted by ↑. The HD is used as a measurement of furthest distance between reference and predicted segmentation. We show the absolute values, where lower is better, as denoted by ↓. The mean results over all IDs can be seen in Table 1.

Table 1. Mean Dice ↑ over the test IDs and Mean HD ↓ over the IDs. Comparison between label loss and KD trained PDD-Nets

Score	Registration		Segmentation
	PDD	PDD_{KD}	nnU-Net
vein Dice %	46.9 ± 4.13	**47.92** ± 4.15	45.93 ± 6.47
artery Dice %	44.48 ± 6.08	*46.67* ± 6.28	**66.80** ± 6.91
overall Dice %	45.69 ± 5.0	*47.3* ± 5.09	**56.36** ± 7.77
vein HD	25.28 ± 166.82	*24.16* ± 159.5	**23.71** ± 366.06
artery HD	28.3 ± 205.19	*27.7* ± 205.54	**26.88** ± 640.51
overall HD	26.79 ± 183.79	*25.93* ± 181.26	**25.33** ± 508.84

We found the distilled network to perform slightly better compared to the label loss trained network over both metrics. When looking at the dice score between the two networks, we found a 2% increase in accuracy over artery segmentation and a 1% increase in vein segmentation. When looking at the HD, we found a similar pattern. The KD trained network outperforms the label loss trained network slightly. We argue that this slight increase is due to the different conceptual representation learned by the distilled network, which would be in line with current research [14,16,21]. The PWC-Net scored at 40.56 ± 3.74 in overall dice and 26.51 ± 160.42 in overall HD on the evaluation videos.

When compared to a 2D segmentation network (nnU-Net [22] Table 1), which was trained on the same image IDs, as the optical flow estimator, we find that the distilled network is performing slightly worse in HD, and worse in Dice score. This result is somewhat expected, since the motion during longer sequences can have significant deformations (compression of veins) and substantial drift. The frame-by-frame segmentation is in principle translation invariant and was trained with a large number of ground truth segmentation annotations. However, when visually looking at estimated segmentations (and quantitatively the variance in HD between the optical flow method and the nnU-Net), we can see that the segmentation network has limited temporal consistency. This suggests that the 2D nnU-Net creates less smooth segmentations over a video, compared to the optical flow method. In the future, we therefore plan to experiment with the optical flow as additional input for a segmentation network. Using a deformation field between two frames, instead of a stacked tensor of all frames, can reduce the computational effort needed for processing, while at the same time containing almost as much information as stacked consecutive frames.

Especially during compression of the vein, this additional information can be leveraged. Figure 4 shows the estimated deformation field between the fixed and the moving frame. The compression is clearly visible as and marked with a black bounding box.

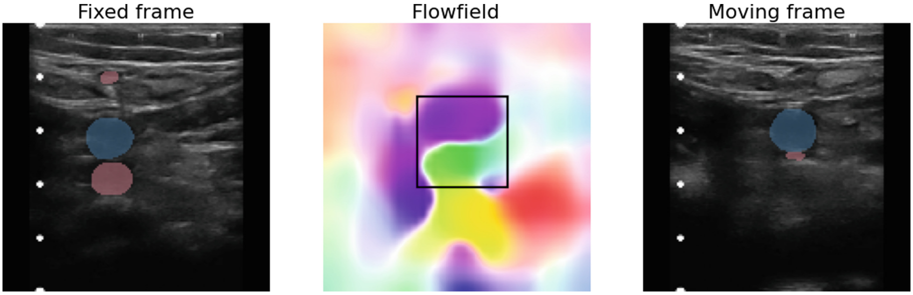


Fig. 4. Visualised deformation field between fixed and moving frame. Segmentation was overlayed for better visibility. The bounding box shows where the compression of the vein (pink) is located and in which direction the vein is compressed. (Color figure online)

4 Conclusion

In this paper, we presented experiments on possible benefits of cross-domain knowledge distillation (from computer vision to medical imaging) for training an optical flow estimator. By using additional teacher-generated soft targets during training, we were able to achieve a small increase in Dice score and a small decrease in Hausdorff distance. This shows that cross-domain KD can have a beneficial effect applied in the training of an image registration network.

We were able to adjust our approach to video inference, such that it is capable of running in realtime, with 1.7 ms per frame pair or more than 500 frames per second. Estimating our approach to use approximately 0.14 GFlops per image, we can calculate an upper limit of roughly 230 frames per second on modern mobile GPUs (Qualcomm Adreno 660).

Performance of segmentation networks still exceeded segmentation via this optical flow based registration of the labels. But we suggest an increase in the segmentation networks' accuracy is possible by combining optical flow information with image features, to add temporal context to the segmentation formation. This was already suggested in previous research in medical video segmentation [23], where improved temporal coherence is reported when optical flow is incorporated. Therefore, we will further investigate the influence of optical flow on vessel segmentation in ultrasound videos.

The results are part of a masters thesis and the code is made available via a git repository[3].

References

1. 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)
2. 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
3. Yao, R., Lin, G., Xia, S., Zhao, J., Zhou, Y.: Video object segmentation and tracking: a survey. ACM Trans. Intell. Syst. Technol. **11** (2020)
4. Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P.: Elastix: a toolbox for intensity-based medical image registration. IEEE Trans. Med. Imaging **29**(1), 196–205 (2009)
5. Mok, T.C.W., Chung, A.C.S.: Conditional deformable image registration with convolutional neural network. In: De Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12904, pp. 35–45. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87202-1_4
6. Hu, Y., et al.: Weakly-supervised convolutional neural networks for multimodal image registration. Med. Image Anal. **49**, 1–13 (2018)
7. Noble, J.A.: Reflections on ultrasound image analysis (2016)
8. De Luca, V., et al.: Evaluation of 2d and 3d ultrasound tracking algorithms and impact on ultrasound-guided liver radiotherapy margins. Med. Phys. **45**(11), 4986–5003 (2018)
9. Liu, F., Liu, D., Tian, J., Xie, X., Yang, X., Wang, K.: Cascaded one-shot deformable convolutional neural networks: developing a deep learning model for respiratory motion estimation in ultrasound sequences. Med. Image Anal. **65**, 101793 (2020)

[3] https://github.com/TillNicke/KD-for-optical-flow.git.

10. Tanno, R.: AutoDVT: joint real-time classification for vein compressibility analysis in deep vein thrombosis ultrasound diagnostics. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11071, pp. 905–912. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00934-2_100
11. Dosovitskiy, A., et al.: Flownet: learning optical flow with convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2758–2766 (2015)
12. Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T.: FlowNet 2.0: evolution of optical flow estimation with deep networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2462–2470 (2017)
13. 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
14. Li, J., Zhao, R., Huang, J.-T., Gong, Y.: Learning small-size DNN with output-distribution-based criteria. In: Fifteenth Annual Conference of the International Speech Communication Association (2014)
15. Yuan, L., Tay, F.E., Li, G., Wang, T., Feng, J. : Revisiting knowledge distillation via label smoothing regularization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3903–3911 (2020)
16. Kim, T., Oh, J., Kim, N., Cho, S., Yun, S.Y.: Comparing kullback-leibler divergence and mean squared error loss in knowledge distillation, arXiv preprint arXiv:2105.08919 (2021)
17. Hering, A., et al.: Learn2Reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning, arXiv preprint arXiv:2112.04489 (2021)
18. Heinrich, M.P., Oktay, O., Bouteldja, N.: Obelisk-net: fewer layers to solve 3d multi-organ segmentation with sparse deformable convolutions. Med. Image Anal. **54**, 1–9 (2019)
19. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected CRFs with gaussian edge potentials. Adv. Neural Inf. Process. Syst. **24**, 109–117 (2011)
20. Reinke, A., et al.: Common limitations of image processing metrics: a picture story, arXiv preprint arXiv:2104.05642 (2021)
21. Hofstätter, S., Althammer, S., Schröder, M., Sertkan, M., Hanbury, A.: Improving efficient neural ranking models with cross-architecture knowledge distillation, arXiv preprint arXiv:2010.02666 (2020)
22. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)
23. Yan, W., Wang, Y., Li, Z., van der Geest, R.J., Tao, Q.: Left ventricle segmentation via optical-flow-net from short-axis cine MRI: preserving the temporal coherence of cardiac motion. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11073, pp. 613–621. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00937-3_70

Metrics/Losses

Motion Correction in Low SNR MRI Using an Approximate Rician Log-Likelihood

Ivor J. A. Simpson[1](✉), Balázs Örzsik[2], Neil Harrison[3], Iris Asllani[2,4], and Mara Cercignani[3]

[1] Department of Informatics, University of Sussex, Brighton, UK
i.simpson@sussex.ac.uk
[2] CISC, Brighton and Sussex Medical School, Brighton, UK
[3] CUBRIC, University of Cardiff, Cardiff, UK
[4] Biomedical Engineering, Rochester Institute of Technology, Rochester, USA

Abstract. Certain MRI acquisitions, such as Sodium imaging, produce data with very low signal-to-noise ratio (SNR). One approach to improve SNR is to acquire several images, each of which takes may take more than a minute, and then average these measurements. A consequence of such a lengthy acquisition procedure is subject motion between each image. This work investigates a solution for retrospective motion correction in this scenario, where the high level of Rician noise renders standard registration tools less effective. We employ a simple generative model for the data based on tissue segmentation maps, and provide a differentiable approximation of the Rician log-likelihood to fit the model to the observations. We find that this approach substantially outperforms a Gaussian log-likelihood baseline on synthetic data that has been corrupted by Rician noise of varying degrees. We also provide results of our approach on real Sodium MRI data, and demonstrate that we can reduce the effects of substantial motion compared to a general purpose registration tool.

Keywords: Motion correction · Rician distribution · Low SNR

1 Introduction

Subject motion is a common issue in long MRI acquisition protocols; in situations where several images have been acquired, motion can be retrospectively corrected using image registration. For brain MRI images with reasonable signal-to-noise-ratio (SNR), general purpose linear image registration tools, e.g. [2,3,8,10,13], have been shown to be highly effective. However, in low SNR MRI data, such as acquired with Sodium MRI, traditional cost functions may become less effective. One cause is the noise properties of the analysed data, which consists of the magnitude of the complex signal components. The noise in such data is described using a Rician distribution [6]. When the SNR of the acquired complex signal is high, the resulting noise is approximately Gaussian. Conversely, when the SNR

A. Hering et al. (Eds.): WBIR 2022, LNCS 13386, pp. 147–155, 2022.
https://doi.org/10.1007/978-3-031-11203-4_16

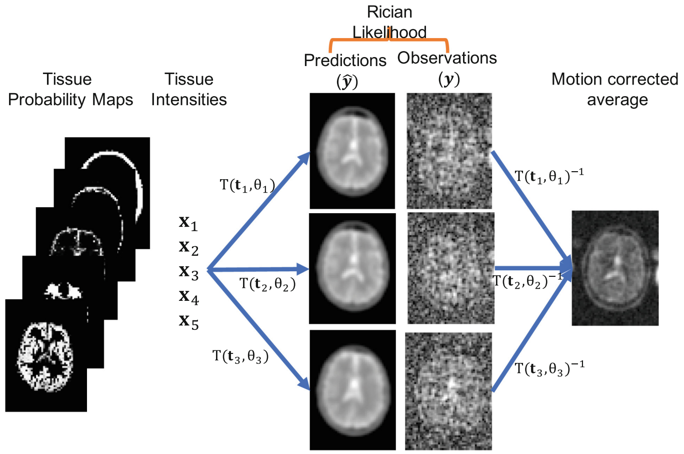

Rician Likelihood
Predictions ($\hat{\boldsymbol{y}}$)
Observations ($\boldsymbol{y}$)
Tissue Probability Maps
Tissue Intensities
Motion corrected average
$\mathbf{x}_1$
$\mathbf{x}_2$
$\mathbf{x}_3$
$\mathbf{x}_4$
$\mathbf{x}_5$
$\mathrm{T}(\mathbf{t}_1,\theta_1)$
$\mathrm{T}(\mathbf{t}_2,\theta_2)$
$\mathrm{T}(\mathbf{t}_3,\theta_3)$
$\mathrm{T}(\mathbf{t}_1,\theta_1)^{-1}$
$\mathrm{T}(\mathbf{t}_2,\theta_2)^{-1}$
$\mathrm{T}(\mathbf{t}_3,\theta_3)^{-1}$

Fig. 1. Illustration of the generative model, which predicts noise free images, $\hat{\mathbf{y}}$, parameterised by: T1 estimated tissue segmentation maps, G, multiplied by estimated tissue intensities, $\mathbf{x}$. These are transformed by a translation, $\mathbf{t}$ and rotation $\boldsymbol{\theta}$. The error between the observations, $\mathbf{y}$ and predictions is described using a Rician likelihood, which is used to drive the parameter estimation.

is low the Rician distribution is asymmetric and dissimilar from a Gaussian. This distinction is particularly significant for registration approaches considering cost functions derived from a Gaussian, e.g. sum-of-squared differences.

This paper introduces a linear motion correction model using a simple generative model of the data. This is inspired by the seminal "Unified Segmentation" paper [1]. A diagram of our approach is given in Fig. 1. Our model produces noise-free predictions, which are rigidly aligned to each of the observed images. The novel contribution of this work lies in our approximation of the Rician log-likelihood that enables gradient estimates through automatic differentiation [14]. This is in contrast to previous work using Rician likelihoods for motion correction [16], which required a gradient-free optimisation of the transformations.

We demonstrate how our approach can be used to remove substantial motion from Sodium MRI data. Sodium is an emerging imaging modality, with several potential biomedical applications [11,19]. However, it has poor SNR due to the relatively low concentration and magnetic susceptibility of Sodium, as shown in Fig. 1. Our results illustrate the effectiveness of this approach in removing substantial motion from high noise situations in both real and synthetic datasets.

2 Background: The Rice Distribution

The noise in magnitude MR images is known to follow a Rice distribution [6]:

$$p(\mathbf{y}|\hat{\mathbf{y}},\sigma) = \mathrm{Rice}(\mathbf{y};\hat{\mathbf{y}},\sigma) = \frac{\mathbf{y}}{\sigma^2}\exp\left(\frac{-(\mathbf{y}^2+\hat{\mathbf{y}}^2)}{2\sigma^2}\right) I_0\left(\frac{\mathbf{y}\hat{\mathbf{y}}}{\sigma^2}\right) \quad (1)$$

where I_0 is a modified Bessel function of the first kind with order zero (described in Sect. 3.2). Unlike the Gaussian, this distribution: is not symmetric with respect to its first parameter, $\hat{\mathbf{y}}$; does not fulfill any of the algebraic conjugacy properties that enable derivation of closed-form parmeter updates; it also does not provide an obvious cost function for directly comparing two images, as it requires a parameterisation in terms of the clean signal, $\hat{\mathbf{y}}$. Generative models can be used to provide such a parameterisation [1].

3 Method

We consider a generative model for the image data based on 5 probabilistic tissue segmentation maps, G, derived from a T1 image acquired in the same space. We denote G as a matrix of size $N \times 5$, where N corresponds to the number of voxels. The intensity of any voxel can be predicted by matrix multiplication with $\mathbf{x}$, a vector containing the intensity for each tissue class. We consider a geometric transformation associated with each observed image:

$$\hat{\mathbf{y}}_i = \mathrm{P}(\mathrm{T}(G\mathbf{x}, \mathbf{t}_i, \boldsymbol{\theta}_i)) \tag{2}$$

where T provides a rigid transformation of $G\mathbf{x}$, according to translation $\mathbf{t}$ and rotation parameters given by $\boldsymbol{\theta}$. We also include a convolution, P, which corresponds to the point-spread function of the acquisition sequence; this is estimated a-priori from the sequence reconstruction method [18]. The predictions $\hat{\mathbf{y}}_i$ can now be fit to the observed data $\mathbf{y}_i$ using an appropriate likelihood function.

3.1 Priors

In this problem, we are considering the registration of noisy data. Accordingly, the model requires the specification of prior knowledge to enable robust inference. We choose a physiologically based Gaussian prior over the concentration of Sodium, measured in mM, for different tissue types:

$$p(\mathbf{x}) = \mathcal{N}([40, 30, 140, 50, 50], [4, 4, 6, 10, 10]^2) \tag{3}$$

where the means are from [11] and the standard deviations are empirically selected.

The translations have a Normal prior, with a standard deviation specified in mm. The rotations, which are described through an axis-angle representation (in Radians), also employ a Normal prior distribution:

$$p(\mathbf{t}_i) = \mathcal{N}(0, 1.25^2)$$
$$p(\boldsymbol{\theta}_i) = \mathcal{N}(0, 0.025^2)$$

3.2 A Stable Approximation of the Rician Log-Likelihood

Most of the Rician likelihood (Eq. 1) is amenable to efficient calculation in a differentiable manner. However, I_0, corresponds to a modified Bessel function of the first kind with order zero [20], which is an infinite series:

$$I_0(z) = \sum_{k=0}^{\infty} \frac{(\frac{1}{4}z^2)^k}{(k!)^2} \tag{4}$$

The result can be approximated as a sum of the first N_k terms. However, this necessitates a differentiable form for the factorial in the denominator. By noting both that $k! = \Gamma(k+1)$, where Γ is the Gamma function, and that we only require the log probability, we can write an approximation for $\log I_0(z)$ as:

$$\log I_0(z) \approx \text{log-sum-exp}(\mathbf{k}(\log(0.25) + 2 * \log(z)) - 2 \ln \Gamma(\mathbf{k}+1)) \tag{5}$$

where $\ln \Gamma$ refers to the log Gamma function. $\mathbf{k}$ is a vector containing values from 0 to N_k, which is summed over. log-sum-exp($\mathbf{z}$) is a numerically stable and convex function [5] for calculating the logarithm of the sum of exponentiated terms, log-sum-exp($\mathbf{z}$) $= \log(\sum_i \exp(z_i))$. This implementation is empirically numerically stable, although inefficient in terms of memory as we require multiplying each voxel by N_k values. We found that $N_k = 50$ provided sufficient precision.

3.3 Inference

We perform maximum-a-posteriori (MAP) inference on the model parameters $\Theta = \{\mathbf{x}, \mathbf{t}, \boldsymbol{\theta}, \sigma\}$, with the following cost function:

$$\mathcal{L} = -\sum_{i}^{N} [\log p(\mathbf{y}_i | \mathbf{x}, \mathbf{t}_i, \boldsymbol{\theta}_i, \sigma) + \log p(\mathbf{t}_i) + \log p(\boldsymbol{\theta}_i)] + \log p(\mathbf{x}) \tag{6}$$

Updates alternated between two groups of parameters, those that are shared for all images $\Theta_1 = \{\mathbf{x}, \sigma\}$ and those that vary per image $\Theta_2 = \{\mathbf{t}, \boldsymbol{\theta}\}$. The updates for Θ_1 were calculated using batches of 5 images at a time, and Θ_2 were updated per image. To account for the batching in updating Θ_1, we perform two update steps on these parameters for every step for Θ_2. The Adam [9] optimiser was used to optimise the model parameters, with a fixed learning rate of $2e^{-2}$ for Θ_1 and $1e^{-3}$ for Θ_2 with $\beta_1 = 0.0$ and $\beta_2 = 0.9$. We stopped the inference after 300 rounds of iterations, at which point the model parameters appeared to have converged. This took approximately 3.5 min for 16 images, or 4.5 min for 32 images on an NVIDIA Quadro RTX 6000 with 24 GB of RAM.

4 Experiments

4.1 Synthetic Data

We generate synthetic data by drawing samples from our generative model with random tissue parameters, drawn from Eq. 3, with additional random voxelwise variability with standard deviations $[4, 4, 6, 10, 10]$ mM. These synthetic images were then transformed to simulate random motion, with translations sampled from $\mathcal{N}(0,\ 5\,\text{mm}^2)$ and angles from $\mathcal{N}(0, 0.1^2)$. Each of these images was then corrupted with Rician noise at various levels. We then tried to correct for the simulated motion using our model with either a Gaussian or Rician likelihood.

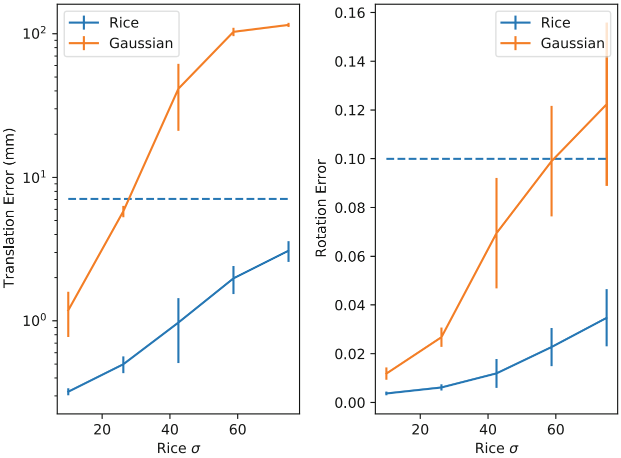


Fig. 2. Synthetic data experiments where the ground truth translation (mean euclidean distance) and rotation error (mean Frobenius norm of the difference of log matrices) are given in the above plots for varying Rician noise level. The dashed line indicates the average initial error. As can be seen, the error when using a Gaussian likelihood rises very quickly, whereas the Rician likelihood is less affected by noise. In this example, $\sigma = 40$ is roughly equivalent to the Sodium MRI data.

Figure 2 illustrates that using the correct likelihood model has a substantial impact on registration performance, particularly in high noise scenarios.

4.2 Real Sodium MRI

^{23}Na MR images were acquired using a dual-tuned, 2-channel (one channel for sodium and one for proton) birdcage ^{23}Na ^{1}H coil developed by RAPID Biomedical GmbH on a 3T Siemens Prisma scanner. Sodium images were acquired using the FLORET spiral sequence [15] with parameters TR $= 120$ ms, TE $= 0.2$ ms, FOV $= 256 \times 256 \times 256$ mm, flip angle $= 80°$, 3 hubs at 22°, 200 interleaves, pulse duration= 0.5ms and dwell time $= 0.01$ ms. Each acquisition took 1 min and 10 s, and was repeated either 16 or 32 times. The k-space data were transferred offline and image reconstruction was performed in Matlab using 3D regridding [15] with density compensation [22]. The data was reconstructed with an isotropic resolution of 4 mm^3 and an image size of $64 \times 64 \times 64$. Examples slices are shown in Fig. 1. A T1-weighted image (2 mm^3 isotropic) was also acquired using the same coil prior to the Sodium data. This was used for preparing tissue segmentation maps using SPM12.

To enable quantification, a set of 4 Sodium phantoms with known concentrations (30, 50, 70, 120 mM) were attached to the head. We use these to map the tissue specific priors, defined in Eq. 3, to the correct intensity range in each image. This mapping is inferred through linear regression of the median signal for each of these phantoms from the true concentrations.

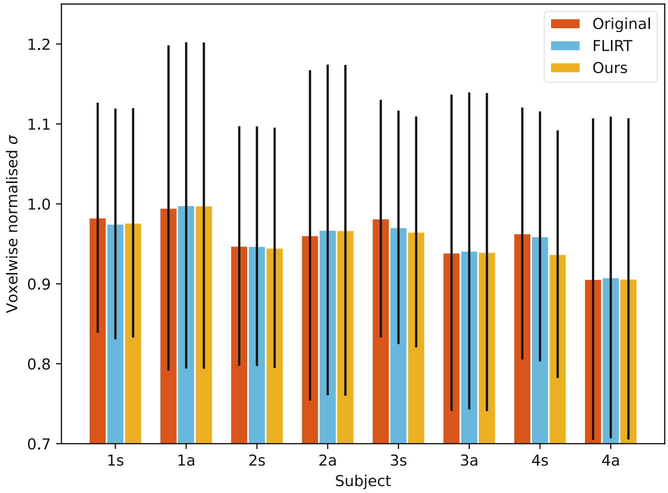


Fig. 3. Bar chart illustrating the mean and std. dev. voxelwise σ, estimated using scipy.stats.Rice, over the motion corrected images for 4 subjects either sleeping (s) or awake (a). These numbers are normalised by the estimated σ in the background.

Using this acquisition protocol, we collected data for 4 subjects either when they are asleep (32 Sodium images) or awake (16 images). The data acquired when sleeping is much more likely to contain motion artefacts due to both the length of the scan and unintentional movements during sleep. Accordingly, we use a more permissive transformation prior (with double the standard deviation for rotation and translations) for these examples.

We experiment with motion correcting the sodium magnitude images using either our proposed approach or “mcflirt” [8], using a cost function of normalized correlation and co-registering to the average image. Nearest neighbour interpolation was used as the final step for both approaches for comparable results without introducing additional smoothness or distortion of noise characteristics.

Validation of the proposed model is complicated by the low SNR exhibited in the motion corrected and averaged images, see Fig. 4 for some examples. Desirable properties of aligned images include: similar values at each voxel over images, and sharp boundaries between regions in the average image. We can measure the first of these by fitting a Rice distribution to each voxel, see Fig. 3.

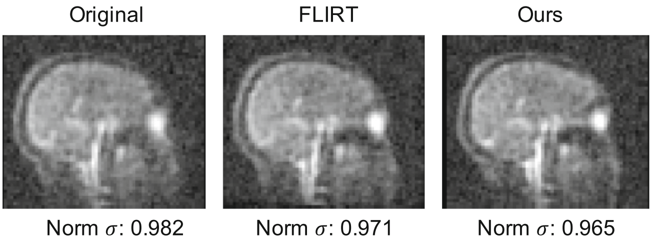


Fig. 4. Example average images calculated by averaging 32 Sodium MRI acquisition. Norm σ refers to the mean voxelwise Rice σ, normalised by an estimate of σ in the image background. In this example, where a lot of motion was detected, our approach leads to a visibly sharper average image.

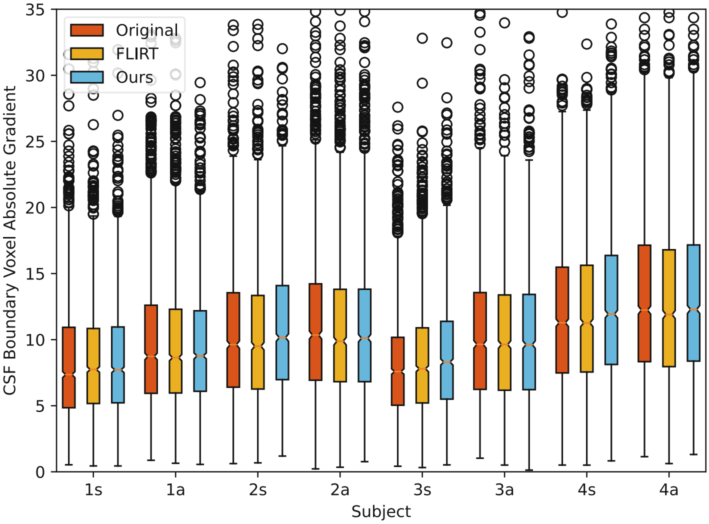


Fig. 5. Boxplot illustrating the absolute gradient of the average image in voxels on the boundary of CSF. Larger values indicate the presence of stronger edges.

We observe that for sleeping acquisitions that are corrupted with visible motion, particularly subject 3 and 4, our approach reduces the mean voxelwise noise compared to other methods. However, in some awake acquisitions, the use of either motion correction approach increases σ; we hypothesise this may be due to interpolation artefacts when correcting sub-voxel motion.

Considering the sharpness of the average image, we can visually observe sharper looking average images in examples with large motion, particularly in subject 3 shown in Fig. 4, where we estimated a mean translation of 7.75 mm (6.49 mm std. dev.) and rotation norm of 0.145 (0.12 std. dev.). To quantify the image sharpness, we examine the distribution of absolute gradient values in voxels that lie on the boundary between CSF and anything else, which should have high contrast. We observe that our motion correction induces stronger edges in most of the acquisitions of sleeping participants.

5 Discussion

The presented approach uses a very simple generative model for the image data, which prevents it overfitting to the high level of noise in the data. However, it also prevents it from making use of strong distinctive features such as the eyes, which contain a high level of Sodium, or the phantoms that are attached to the head. Future work will consider using more complex statistical models and techniques, such as variational inference [7], to build a voxelwise generative model. Amortised inference strategies could also be investigated to improve efficiency [4].

In our experimentation, we observed that in some cases where low motion was observed, our algorithm overestimated the level of movement. We found that this was removed by introducing variable transformation permissiveness based on our prior beliefs on the level of motion. Future work will consider methods for inferring these parameters, and using auto-regressive priors on motion [21].

This work has not investigated preprocessing the data using denoising methods, e.g. [12]; although such approaches may produce cleaner representations for aligning the data, they also manipulate the underlying image statistics being modelled, which may lead to biased results. We also have not compared against the use of robust cost functions [17], although these are generally more suited to heavy tailed rather than asymmetric noise distributions as we have here.

We have published our code on GitHub[1]. The data are not currently available for distribution as the initial analysis of a wider dataset is ongoing.

6 Conclusions

This paper has introduced an algorithm for data modelling and motion correction of low SNR MRI data using a differentiable approximation of the Rician log-likelihood. Our synthetic experiments illustrated the importance of choosing the right cost function for generative models for motion correction, as the Gaussian likelihood performs very poorly where the errors take a different form. On real Sodium MRI data, our results provide support for the use of our method in resolving substantial motion artefacts and creating sharper average images.

Acknowledgments. We acknowledge funding from the University of Sussex used in the data acquisition. We thank Guillaume Madelin for providing the Sodium MRI sequence.

References

1. Ashburner, J., Friston, K.J.: Unified segmentation. Neuroimage **26**(3), 839–851 (2005)
2. Ashburner, J., Neelin, P., Collins, D., Evans, A., Friston, K.: Incorporating prior knowledge into image registration. Neuroimage **6**(4), 344–352 (1997)

[1] https://github.com/ivorsimpson/sodium-mri-inference.

3. Avants, B.B., Tustison, N., Song, G., et al.: Advanced normalization tools (ants). Insight j **2**(365), 1–35 (2009)
4. 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)
5. Boyd, S., Boyd, S.P., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)
6. Gudbjartsson, H., Patz, S.: The Rician distribution of noisy MRI data. Magn. Reson. Med. **34**(6), 910–914 (1995)
7. Hoffman, M.D., Blei, D.M., Wang, C., Paisley, J.: Stochastic variational inference. J. Mach. Learn. Res. **14**(40), 1303–1347 (2013). https://jmlr.org/papers/v14/hoffman13a.bib
8. 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)
9. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
10. Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P.: Elastix: a toolbox for intensity-based medical image registration. IEEE Trans. Med. Imaging **29**(1), 196–205 (2009)
11. Madelin, G., Regatte, R.R.: Biomedical applications of sodium MRI in vivo. J. Magn. Reson. Imaging **38**(3), 511–529 (2013)
12. Manjón, J.V., Carbonell-Caballero, J., Lull, J.J., García-Martí, G., Martí-Bonmatí, L., Robles, M.: MRI denoising using non-local means. Med. Image Anal. **12**(4), 514–523 (2008)
13. Ourselin, S., Roche, A., Subsol, G., Pennec, X., Ayache, N.: Reconstructing a 3D structure from serial histological sections. Image Vis. Comput. **19**(1–2), 25–31 (2001)
14. Paszke, A., et al.: Automatic differentiation in PyTorch (2017)
15. Pipe, J.G., Zwart, N.R., Aboussouan, E.A., Robison, R.K., Devaraj, A., Johnson, K.O.: A new design and rationale for 3D orthogonally oversampled k-space trajectories. Magn. Reson. Med. **66**(5), 1303–1311 (2011)
16. Ramos-Llordén, G., Arnold, J., Van Steenkiste, G., Van Audekerke, J., Verhoye, M., Sijbers, J.: Simultaneous motion correction and t1 estimation in quantitative t1 mapping: an ml restoration approach. In: 2015 IEEE International Conference on Image Processing (ICIP), pp. 3160–3164. IEEE (2015)
17. Reuter, M., Rosas, H.D., Fischl, B.: Highly accurate inverse consistent registration: a robust approach. Neuroimage **53**(4), 1181–1196 (2010)
18. Riemer, F., Solanky, B.S., Stehning, C., Clemence, M., Wheeler-Kingshott, C.A., Golay, X.: Sodium (23Na) ultra-short echo time imaging in the human brain using a 3D-cones trajectory. Magn. Reson. Mater. Phys., Biol. Med. **27**(1), 35–46 (2014)
19. Rose, A.M., Valdes, R., Jr.: Understanding the sodium pump and its relevance to disease. Clin. Chem. **40**(9), 1674–1685 (1994)
20. Wolfram Mathworld: Modified Bessel function of the first kind. https://mathworld.wolfram.com/ModifiedBesselFunctionoftheFirstKind.html
21. Woolrich, M.W., Jenkinson, M., Brady, J.M., Smith, S.M.: Fully Bayesian spatio-temporal modeling of fMRI data. IEEE Trans. Med. Imaging **23**(2), 213–231 (2004)
22. Zwart, N.R., Johnson, K.O., Pipe, J.G.: Efficient sample density estimation by combining gridding and an optimized kernel. Magn. Reson. Med. **67**(3), 701–710 (2012)

Cross-Sim-NGF: FFT-Based Global Rigid Multimodal Alignment of Image Volumes Using Normalized Gradient Fields

Johan Öfverstedt(✉), Joakim Lindblad, and Nataša Sladoje

Department of Information Technology, Uppsala University, Uppsala, Sweden
johan.ofverstedt@it.uu.se

Abstract. Multimodal image alignment involves finding spatial correspondences between volumes varying in appearance and structure. Automated alignment methods are often based on local optimization that can be highly sensitive to initialization. We propose a novel efficient algorithm for computing similarity of normalized gradient fields (NGF) in the frequency domain, which we globally optimize to achieve rigid multimodal 3D image alignment. We validate the method experimentally on a dataset comprised of 20 brain volumes acquired in four modalities (T1w, Flair, CT, [18F] FDG PET), synthetically displaced with known transformations. The proposed method exhibits excellent performance on all six possible modality combinations and outperforms the four considered reference methods by a large margin. An important advantage of the method is its speed; global rigid alignment of 3.4 Mvoxel volumes requires approximately 40 s of computation, and the proposed algorithm outperforms a direct algorithm for the same task by more than three orders of magnitude. Open-source code is provided.

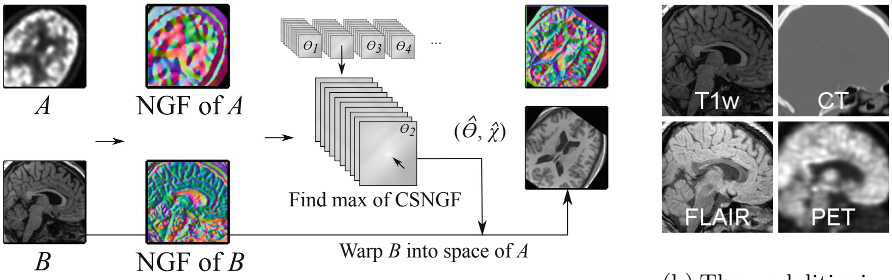
Keywords: Image registration · Global · Exhaustive search · NGF · FFT · Matching · GPU implementation

1 Introduction

Multimodal image alignment (also known as registration) involves finding correspondences between images with varying degrees of difference of appearance and structure [18], often applied with the goal of combining the complementary information of each modality via image fusion. Alignment of large displacements is particularly challenging since correspondences to be inferred are far apart and presence of multiple local optima becomes increasingly problematic as the search space grows, thereby often requiring global contextual and spatial information.

A large number of methods exist for multimodal alignment [14], including local optimization methods based on mutual information (MI) [8,17] or normalized gradient fields (NGF) [3,13], and representation extraction techniques

A. Hering et al. (Eds.): WBIR 2022, LNCS 13386, pp. 156–165, 2022.
https://doi.org/10.1007/978-3-031-11203-4_17



(a) Illustration of the global image volume alignment method.

(b) The modalities included in this study.

Fig. 1. Main steps of one level of the multi-level rigid alignment method (a), and examples of the modalities considered in the evaluation (b) (images from [9]). (a) Two image volumes of modalities A (here [18F] FDG PET), and B (T1 weighted MR), are used as input. For a set of 3D rotations $\boldsymbol{\theta}$, the similarity measure s_{ANGF} between the NGF of A and the NGF of B (rotated), (here shown as RGB images where each color channel represents one component of the 3D vector field $n_1(\cdot; A), n_2(\cdot; A), n_3(\cdot; A)$) is computed for all 3D displacements. The rigid alignment $(\hat{\boldsymbol{\theta}}, \hat{\chi})$ is found by locating the maximum s_{ANGF}.

based on local self-similarities [4] or Deep Feature Learning [5,12]. Most of the (intensity-based) methods are based on some form of local optimization, which usually require a good initial guess to work well. However, several global alignment methods do exist, including [1,6] as well as a recently proposed method based on the cross-mutual information function (CMIF) [10].

We propose a new global alignment method based on NGF that is fast and exhibits excellent performance on a rigid multimodal 3D medical image alignment task. Our evaluation on 6 pairs of modality combinations shows that it outperforms well known methods which rely on local optimization of MI [8,17] and NGF [3] as well as the recently proposed approach based on global optimization of CMIF [10]. Figure 1 illustrates the general idea of the method.

A fast PyTorch-based implementation of the method is shared as open-source at http://github.com/MIDA-group/cross_sim_ngf.

2 Background

The (regularized) normalized gradient field [3], for image A at point x, is

$$\boldsymbol{n}(x; A) = \frac{\nabla A(x)}{\sqrt{\|\nabla A(x)\|_2^2 + \epsilon^2}}, \tag{1}$$

where ϵ is a small constant to reduce the impact of gradients with very small magnitude and avoid division by zero. In this work we use $\epsilon = 10^{-5}$ for $A(x) \in [0, 1]$, selected empirically; higher values yielded more failed alignments and lower values mostly made the measure more noisy.

The main assumption of NGF-based alignment is that parts of images (acquired by different modalities) are in correspondence when the directions of their intensity changes are parallel or anti-parallel. A local similarity of NGF (SNGF) based on the squared dot-product of the elements of the NGF is defined as

$$s_{\mathrm{NGF}}(x; A, B) = \langle \boldsymbol{n}(x; A), \boldsymbol{n}(x; B) \rangle^2 . \tag{2}$$

Orientation correlation (OC) and squared orientation correlation (SOC) offer an efficient way of computing SNGF of 2D images for all discrete displacements [1]. In 2D, the vectors $\boldsymbol{n}(\cdot; \cdot)$ are represented as complex numbers. A fast algorithm utilizing log-polar Fourier transform for OC-based alignment w.r.t.rotation and scaling is proposed in [16]. A computationally efficient extension to 3D [2] required a modification of the similarity measure; the authors proposed to, instead of (2), use its unsquared version:

$$s_{\text{US-NGF}}(x; A, B) = \langle \boldsymbol{n}(x; A), \boldsymbol{n}(x; B) \rangle . \tag{3}$$

By observing three separable components of the (unsquared) dot-product in (3), the authors [2] formulated an algorithm for efficiently computing the measure for all discrete displacements using cross-correlation in the frequency domain. None of the existing work, however, describes a method for computing similarities of NGF using the squared measure (2) efficiently in the frequency domain for 3D volumes, a gap which we aim to fill with this work.

The ability to use the squared measure rather than the unsquared measure is beneficial for multimodal image alignment [1]. Equation (3), similarly to (the unsquared) OC [1], exhibits useful properties such as invariance to changes of contrast and absolute intensity levels, which are suitable for monomodal registration tasks. However, multimodal scenarios are often characterized by the appearance of parts of a sample that are dark in one modality and bright in another; in such cases, aligned samples actually minimize $s_{\text{US-NGF}}$.

3 Method

Here we define a similarity measure between NGF based on (2), a cross-similarity (*c.f.*cross-correlation) formulation of the measure, and propose an algorithm for computing it efficiently in the frequency domain for all 3D discrete displacements.

In [3], the point-wise contributions of s_{NGF} (2) are aggregated by summation. A downside of this choice is that it imposes a strong bias towards full overlap of the images which can be especially problematic for global optimization. We instead formulate a scaled similarity measure that is applied to selected regions of the images $A\colon X_A \to \mathbb{R}$ and $B\colon X_B \to \mathbb{R}$, defined by indicator functions (masks) $M_A\colon X_A \to \{0,1\}$ and $M_B\colon X_B \to \{0,1\}$, ignoring the parts of the finite rectangular domains where either M_A or M_B are zero-valued. The average similarity of NGF is

$$s_{\mathrm{ANGF}}(A, B; M_A, M_B) = \frac{1}{\sum_x M_A(x) M_B(x)} \sum_x M_A(x) M_B(x) s_{\mathrm{NGF}}(x; A, B) . \tag{4}$$

Based on s_{ANGF}, we define the *Cross Similarity of NGF*

$$\text{CSNGF}(\chi; A, B, M_A, M_B) = \frac{1}{N(\chi)} \sum_x M_A(x) M_B(x+\chi) s_{\text{NGF}}(x; A(x), B(x+\chi)), \tag{5}$$

where $\chi \in S$ is a discrete translation from the set S representing all the considered discrete translations and $N(\chi)$ is the number of overlapping voxels (where M_A and M_B intersect) as a function of χ. $N(\chi)$ can be computed as the cross-correlation between the mask images $N(\chi) = (M_A \star M_B)(\chi)$. An analogous approach is taken in [10] to compute CMIF. Masks are essential for computation of CSNGF, for any choice of S which results in a partial overlap of the images. Figure 1 illustrates CSNGF as a part of a rigid 3D alignment method.

A *direct method* for computing CSNGF for all $\chi \in S$ involves looping over each χ, and compute and aggregate s_{NGF} for all overlapping voxels. If $|S| = O(|X_A|)$, then the run-time complexity of the direct method is $O(|X_A||X_B|)$ which for equisized images A and B gives a quadratic run-time complexity in the size of the images, which is not feasible for volumes of realistic sizes.

We propose a more efficient algorithm for computing CSNGF for all $\chi \in S$ in 3D. By reformulating (2), and expanding the squared dot-product,

$$s_{\text{NGF}}(x; A, B) = \sum_{i=1}^{3} \Big(\boldsymbol{n}_i(x;A)^2 \boldsymbol{n}_i(x;B)^2 + 2 \sum_{j=i+1}^{3} \boldsymbol{n}_i(x;A) \boldsymbol{n}_j(x;A) \boldsymbol{n}_i(x;B) \boldsymbol{n}_j(x;B) \Big), \tag{6}$$

we express it as 6 separable parts comprising 3 squared components ($i \in \{1, 2, 3\}$), as well as products of 3 pairs of components ($(i, j) \in \{(1, 2), (1, 3), (2, 3)\}$), of the NGF vector fields (see Fig. 1a), which can be computed independently for all χ using cross-correlation. Let $\boldsymbol{n}_i^M$ denote a modified NGF scaled by the associated mask, $\boldsymbol{n}_i^M(x\,; A) = M_A(x)\, \boldsymbol{n}_i(x; A)$. The required cross-correlations $((\boldsymbol{n}_i^M(\cdot\,; A)^2) \star (\boldsymbol{n}_i^M(\cdot\,; B)^2))$ and $((\boldsymbol{n}_i^M(\cdot\,; A)\boldsymbol{n}_j^M(\cdot\,; A)) \star (\boldsymbol{n}_i^M(\cdot\,; B)\boldsymbol{n}_j^M(\cdot\,; B)))$ are efficiently computed in the frequency domain; $(\boldsymbol{n}_i^M(\cdot\,; A)^2 \star \boldsymbol{n}_i^M(\cdot\,; B)^2) = F^{-1}\big(\overline{F(\boldsymbol{n}_i^M(\cdot\,; A)^2)} \odot F(\boldsymbol{n}_i^M(\cdot\,; B)^2)\big)$, where $F(\cdot)$ denotes the Fourier transform, $\overline{z}$ denotes complex conjugation and $\odot$ denotes element-wise multiplication. For efficiency, the 6 separable parts are aggregated in the Fourier domain. Computing CSNGF involves 14 real-valued FFTs (6 per image plus 1 mask per image) and 2 inverse FFTs. Generalization to nD is straightforward.

3.1 Method for Global 3D Rigid Alignment

The fast algorithm for computing CSNGF for all $\chi \in S$ provides direct means of global optimization of s_{ANGF} w.r.t.axis-aligned shifts. To reach global optimization w.r.t.rigid transformations, we adopt a hybrid approach where the space of 3D rotations $\boldsymbol{\theta} = (\theta_x, \theta_y, \theta_z)$ (represented as Euler angles) is explored via a multi-stage combination of Gaussian pyramids, random search, and global optimization of s_{ANGF}. One stage of this coarse-to-fine method is illustrated in Fig. 1. This multi-stage approach facilitates global search at the lowest considered resolution, followed by more local search to refine the alignment.

Initially, a Gaussian resolution pyramid with m levels is constructed through the application of Gaussian blur and downsampling. For each level $k \in \{1 \dots m\}$, a random search is performed in a coarse-to-fine sequence, by sampling angles $\boldsymbol{\theta}$ either (a) as random rotations from the set of all possible rotations, for the first level ($k = 1$), or (b) as rotations close to one of the p_{k-1} best solutions of the previous level, for levels $k \in \{2, \dots, m\}$. An angle "close to" is realized by perturbing the previous solution by a change in rotation around axes (x, y, z), sampled from $\mathbb{U}(-u_{k-1}, u_{k-1})$ for each axis. For each $\boldsymbol{\theta}$, the corresponding transformation $T_{\boldsymbol{\theta}}$ is applied to the floating image $B_{\boldsymbol{\theta}} = B \circ T_{\boldsymbol{\theta}}$ using trilinear interpolation and its mask $M_{B_{\boldsymbol{\theta}}} = M_B \circ T_{\boldsymbol{\theta}}$ using nearest neighbor interpolation. $\boldsymbol{n}(\cdot\,; B_{\boldsymbol{\theta}})$ is computed, followed by computation of $\arg\max_{\chi} \text{CSNGF}(\chi; A, B_{\boldsymbol{\theta}}, M_A, M_{B_{\boldsymbol{\theta}}})$ for all $\chi \in S$, where S is the set of displacements satisfying a user-selected amount of minimum overlap γ. A suitable zero padding scheme is used to enable partial overlaps (following [10]). For $k > 1$, the p_{k-1} best solutions of the previous level are also evaluated unmodified to not risk discarding good solutions. For $k = m$, the best rotation and displacement are taken as the final rigid transformation.

The method is parameterized by blur-levels $\sigma = (\sigma_1, \dots, \sigma_m)$, downsampling factors $\mathbf{d} = (d_1, \dots d_m)$, largest allowed steps $\mathbf{u} = (u_1, \dots u_{m-1})$, number of rotations $\mathbf{a} = (\mathrm{a}_1, \dots \mathrm{a}_m)$, and number of kept best solutions $\mathbf{p} = (p_1, \dots, p_{m-1})$. For all related experiments, $\mathbf{d} = (4, 2, 2, 1)$, $\mathbf{a} = (5000, 3000, 300, 0)$, $\mathbf{u} = (10, 3, 0)$, and $\mathbf{p} = (20, 3, 1)$. We use $\gamma = 0.5$ everywhere in this study.

4 Performance Analysis

The empirical evaluation of the proposed method is based on the CERMEP-IDB-MRXFDG dataset [9], available upon request from the authors. The dataset consists of images of brains of 33 subjects acquired by 4 different modalities: T1 weighted MR, Flair MRI, Computed Tomography (CT), [18F] FDG PET, all mapped to the standard MNI space (see Fig. 1b), thus providing ground-truth for image alignment method evaluation, and a possibility to consider 6 different combinations of modalities, enabling evaluation of the generality of the methods.

4.1 Similarity Landscape of the Average SNGF

First, we perform an empirical analysis of how (4) is affected by spatial transformations of the observed images. The aim is to provide evidence of the relevance of global optimization for multimodal image alignment. We consider two images acquired with the modalities FLAIR and PET and study the similarity landscape as the PET volume is rotated around a single axis of rotation; the result is shown in Fig. 2. We observe that the similarity landscape exhibits characteristics that impede local methods without a good initial guess for all parameters.

4.2 Multimodal Brain Image Volume Alignment

We compare the proposed method with two global and two local alignment methods on the task of recovering rigid transformations of brain image volumes.

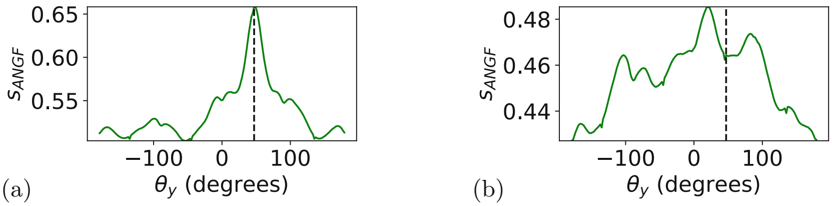
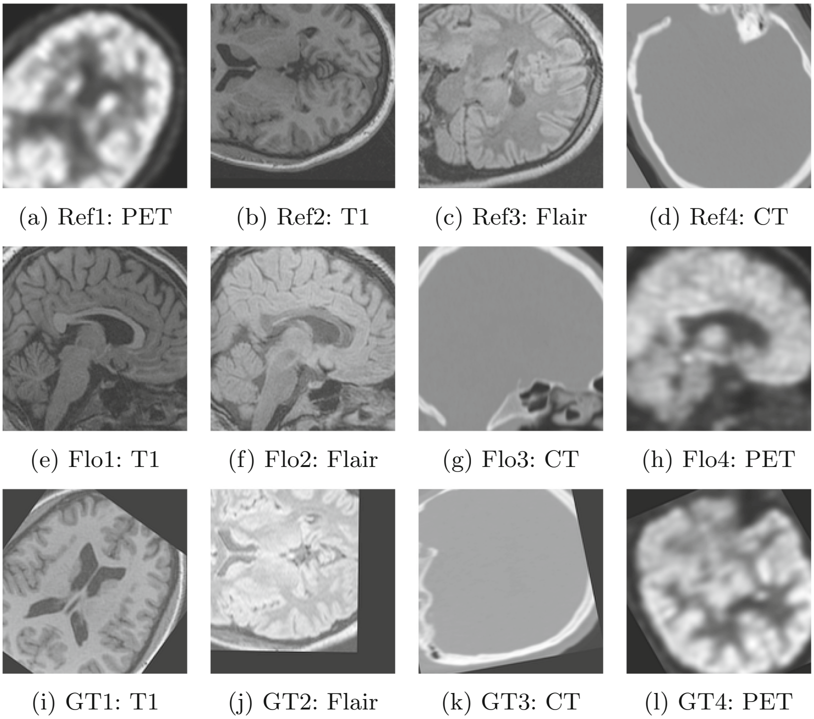


Fig. 2. Similarity landscape of s_{ANGF} for a pair of FLAIR and PET images of a brain (blur: $\sigma = 5$), w.r.t.the rotation angle θ_y. Two scenarios are presented: (a) with no additional transformation, *i.e.*, all transformation parameters other than θ_y have their correct values, and (b) when the FLAIR image has been rotated by 5° around a random axis (other than y) and translated by 20 *vx* in a randomly direction. The vertical dashed lines mark the sought angle. We observe that, (a) even without displacement, the convergence region of the sought angle has a limited size, with local maxima near the global maximum, and that (b) displacements along multiple dimensions make the search using local approaches further challenging; here the sought angle (dashed line) is between local optima.

For each of the twelve (ordered) pairs of modalities (six unordered modality combinations) included in the CERMEP-IDB-MRXFDG dataset, and for each of the first 20 subjects (the last 13 used for parameter tuning), we randomly (uniformly) sample a 3D rotation $\boldsymbol{\theta}$, and an axis-aligned shift $\chi_i \in [-30\, vx, +30\, vx]$ for each axis i. These transformations are applied, using inverse mapping and bicubic interpolation, to the first image volume of each pair. The transformed image is taken as reference image and the untransformed image as floating image in the alignment task. Finally, a block of size 151 × 151 × 151 *vx* (*c.f.*original size 207 × 243 × 226) at the center of the volume is extracted, retaining most of the content of interest, while omitting most of the background and avoiding padding introduced by inverse mapping outside the image domain. This setup enables evaluation of the accuracy of the proposed method w.r.t.alignment of multimodal 3D images by recovering these known transformations. Example slices of pairs from the selected modality combination are shown in Fig. 3.

With the aim to evaluate the benefit of the proposed algorithm, based on the original similarity of NGF (2), compared to the one proposed in [2], we let USNGF refer to an alignment method similar to CSNGF, but with s_{NGF} in (5) replaced by $s_{\text{US-NGF}}$. We evaluate both USNGF and “USNGF-”, where the latter denotes USNGF but with an intensity-inverted floating image, to observe the sensitivity of USNGF to the sign of the gradients [1]. We also include the recently proposed CMIF-based global alignment method [10], which has exhibited excellent performance and outperformed several recent Deep Learning methods (including [12]) on multiple biomedical datasets. The selected global optimization methods are implemented in Python/PyTorch [11] with CUDA/GPU-acceleration.



(a) Ref1: PET (b) Ref2: T1 (c) Ref3: Flair (d) Ref4: CT

(e) Flo1: T1 (f) Flo2: Flair (g) Flo3: CT (h) Flo4: PET

(i) GT1: T1 (j) GT2: Flair (k) GT3: CT (l) GT4: PET

Fig. 3. Sample slices of 3D image pairs from the evaluation dataset generated from the CERMEP-IDB-MRXFDG dataset [9]. (a-d) the reference (transformed) images and (e-h) the floating images. Image (e) is to be registered to (a); (f) to (b), (g) to (c) and (h) to (d). The bottom row shows the ground-truth (GT) of each floating (Flo) image aligned to the corresponding reference (Ref) image.

We also compare with local optimization-based methods using MI and NGF as objective functions, relying on open-source implementations Elastix [8] and AIRLab [15] respectively.

We use the mean Euclidean distance between the corresponding corner points of the extracted block before and after the performed (recovered) alignment as a displacement measure, denoted d_E. We consider an alignment successful if $d_E < 5\,vx$, which is approximately 2% of the length of the diagonal of the blocks.

For CMIF we use $k = 16$ (for the k-means clustering), and $\sigma = (3.0, 1.5, 1.0, 0.0)$. For NGF, USNGF (and USNGF-), we use $\sigma = (5.0, 3.0, 2.0, 1.5)$. For local optimization MI (LO-MI) [8,17], we use 6 pyramid levels, the Adaptive Stochastic Gradient Descent optimizer [7], 4096 maximum iterations for each level. For local optimization NGF (LO-NGF) [3], we use 5 pyramid levels, ADAM optimizer, iteration counts according to the schedule (4096, 4096, 1024, 100, 50), with downsampling factors (16, 8, 4, 2, 1) and Gaussian smoothing parameters (15.0, 9.0, 5.0, 3.0, 1.0), with learning-rate 0.01. Trilinear interpolation is used.

Results. The results of the evaluation of the 6 considered methods on the multimodal brain image dataset are presented in Table 1. The proposed method provides overall excellent performance, and is the best choice for all observed modality combinations. Most of the competitors show generally poor performance, completely failing on one or more modality combinations. Near-successes are also of interest, since those solutions may be refined with a local optimization method; therefore, we plot the distribution up to the threshold $d_E < 20$ as Fig. 4.

Table 1. Image alignment performance presented in terms of success-rate, where the threshold of success is set to 5 *vx*. The modality names are abbreviated in the headings (T: T1, F: Flair, C: CT, P: [18F] FDG PET).

Method	Modalities					
	T/F	T/C	T/P	F/C	F/P	C/P
LO-MI	0.05	0.025	0.075	0.025	0.1	0.075
LO-NGF	0.025	0.00	0.00	0.00	0.00	0.00
CMIF	0.675	0.30	0.325	0.80	0.85	0.525
USNGF	0.225	0.00	0.00	0.00	**0.925**	0.10
USNGF-	0.00	0.275	0.00	0.00	0.00	0.00
CSNGF	**1.00**	**0.95**	**0.925**	**0.90**	**0.925**	**0.95**

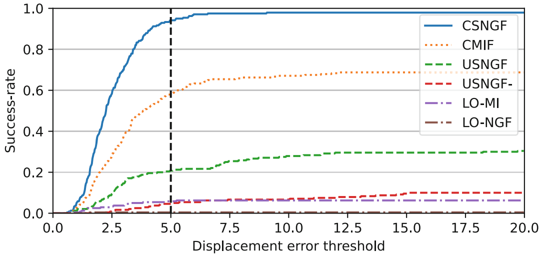


Fig. 4. Success-rate of each considered method as a function of the acceptable displacement error t (fraction of the 240 alignments where $d_E < t$); the results for all modality combinations are aggregated. Up and to the left is better.

4.3 Time Analysis

We compare the run-times of the global rigid registration methods, as well as the run-times of the novel Cross-Sim-NGF algorithm with a direct (not FFT-based) approach. The reported results are obtained on a Nvidia GeForce RTX 3090.

Both the FFT-based algorithm and the direct method are implemented in Python/PyTorch using GPU-acceleration; the direct method consists of a loop over all axis-aligned shifts $\chi \in S$, and computation of the squared dot-products.

The average run-times of the methods CMIF, USNGF, and CSNGF are 569 s, 33 s, and 41 s, respectively. Comparison of the run-times of the FFT-based algorithm and the direct method, as a function of image size, is presented in Table 2. We observe that for size 128, the here proposed algorithm is approximately 6275 times (more than three orders of magnitude) faster.

Table 2. Run-time (s) comparison of FFT-based CSNGF and a direct algorithm for computing CSNGF, for all $\chi \in S$ where the overlap is 50% or higher, on cube image volumes of increasing size (expressed as side-length).

Method	Size				
	8	16	32	64	128
Direct algorithm	0.129	0.557	3.537	27.07	502.4
FFT-based alg.	**0.002**	**0.002**	**0.002**	**0.008**	**0.088**

5 Conclusion

We propose a novel NGF-based method for global rigid 3D multimodal alignment, which extends a well-performing method for 2D image alignment, outperforming a previous extension that relies on an unsquared version of the similarity measure. We confirm both its great performance and its high efficiency. Through the comparison with CMIF-based alignment [10], the method is indirectly compared with several approaches based on deep learning while leaving a more comprehensive comparative study as future work. The method does not use any training (data), which is a large advantage for (bio)medical applications [5].

Acknowledgments. We thank Ines Mérida and team for the CERMEP-IDB-MRXFDG dataset. We acknowledge financial support by Vinnova, MedTech4Health Grant 2017-02447.

References

1. Fitch, A., Kadyrov, A., Christmas, W., Kittler, J.: Orientation correlation. In: Proceedings of British Machine Vision Conference, pp. 133–142 (2002)
2. Fotin, S.V., et al.: Normalized gradient fields cross-correlation for automated detection of prostate in magnetic resonance images. In: Medical Imaging 2012: Image Processing, vol. 8314, p. 83140V. International Society for Optics and Photonics (2012)
3. Haber, E., Modersitzki, J.: Intensity gradient based registration and fusion of multi-modal images. In: Larsen, R., Nielsen, M., Sporring, J. (eds.) MICCAI 2006. LNCS, vol. 4191, pp. 726–733. Springer, Heidelberg (2006). https://doi.org/10.1007/11866763_89
4. Heinrich, M.P., et al.: MIND: modality independent neighbourhood descriptor for multi-modal deformable registration. Med. Image Anal. **16**(7), 1423–1435 (2012)

5. Islam, K.T., Wijewickrema, S., O'Leary, S.: A deep learning based framework for the registration of three dimensional multi-modal medical images of the head. Sci. Rep. **11**(1), 1–13 (2021)
6. Jenkinson, M., Smith, S.: A global optimisation method for robust affine registration of brain images. Med. Image Anal. **5**(2), 143–156 (2001)
7. Klein, S., Pluim, J.P.W., Staring, M., Viergever, M.A.: Adaptive stochastic gradient descent optimisation for image registration. Int. J. Comput. Vis. **81**(3), 227 (2008)
8. Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P.W.: Elastix: a toolbox for intensity-based medical image registration. IEEE Trans. Med. Imaging **29**(1), 196–205 (2010)
9. Mérida, I., et al.: CERMEP-IDB-MRXFDG: a database of 37 normal adult human brain [18F] FDG PET, T1 and FLAIR MRI, and CT images available for research. EJNMMI Res. **11**(1), 1–10 (2021)
10. Öfverstedt, J., Lindblad, J., Sladoje, N.: Fast computation of mutual information in the frequency domain with applications to global multimodal image alignment. arXiv preprint arXiv:2106.14699 (2021)
11. Paszke, A., Gross, S., et al.: PyTorch: an imperative style, high-performance deep learning library. Adv. Neural Inf. Process. Syst. **32**, 8026–8037 (2019)
12. Pielawski, N., et al.: CoMIR: contrastive multimodal image representation for registration. In: Neural Information Processing System, vol. 33, pp. 18433–18444 (2020)
13. Pluim, J., Maintz, J., Viergever, M.: Image registration by maximization of combined mutual information and gradient information. IEEE Trans. Med. Imaging **19**(8), 809–814 (2000)
14. Saiti, E., Theoharis, T.: An application independent review of multimodal 3D registration methods. Comput. Graph. **91**, 153–178 (2020)
15. Sandkühler, R., Jud, C., Andermatt, S., Cattin, P.C.: Airlab: autograd image registration laboratory. arXiv preprint arXiv:1806.09907 (2018)
16. Tzimiropoulos, G., Argyriou, V., Zafeiriou, S., Stathaki, T.: Robust FFT-based scale-invariant image registration with image gradients. IEEE Trans. Pattern Anal. Mach. Intell. **32**(10), 1899–1906 (2010)
17. Viola, P., Wells, W.M., III.: Alignment by maximization of mutual information. Int. J. Comput. Vis. **24**(2), 137–154 (1997)
18. Zitova, B., Flusser, J.: Image registration methods: a survey. Image Vis. Comput. **21**(11), 977–1000 (2003)

Identifying Partial Mouse Brain Microscopy Images from the Allen Reference Atlas Using a Contrastively Learned Semantic Space

Justinas Antanavicius[1], Roberto Leiras[2], and Raghavendra Selvan[1,2(✉)]

[1] Department of Computer Science, University of Copenhagen, Copenhagen, Denmark
raghav@di.ku.dk
[2] Department of Neuroscience, University of Copenhagen, Copenhagen, Denmark

Abstract. Registering mouse brain microscopy images to a reference atlas is crucial to determine the locations of anatomical structures in the brain, which is an essential step for understanding the function of brain circuits. Most existing registration pipelines assume the identity of the reference plate – to which the image slice is to be registered – is known beforehand. This might not always be the case due to three main challenges in microscopy image data: missing image regions (partial data), different cutting angles compared to the atlas plates and a large number of high-resolution images to be identified. Manual identification of reference plates as an initial step requires highly experienced personnel and can be biased, tedious and resource intensive. On the other hand, registering images to all atlas plates can be slow, limiting the application of automated registration methods when dealing with high-resolution image data. This work proposes to perform the image identification by learning a *low-dimensional* space that captures the similarity between microscopy images and the reference atlas plates. We employ Convolutional Neural Networks (CNNs), in the *Siamese* network configuration, to first obtain low-dimensional embeddings of microscopy image data and atlas plates. These embeddings are contrasted with positive and negative examples in order to learn a semantically meaningful space that can be used for identifying corresponding 2D atlas plates. At inference, atlas plates that are closest to the microscopy image data in the learned embedding space are presented as candidates for registration. Our method achieved TOP-3 and TOP-5 accuracy of 83.3% and 100%, respectively, compared to the SimpleElastix-based baseline which obtained 25% in both the Top-3 and Top-5 accuracy (Source code is available at https://github.com/Justinas256/2d-mouse-brain-identification).

Keywords: Image registration · Mouse brain · Partial data · Deep learning

A. Hering et al. (Eds.): WBIR 2022, LNCS 13386, pp. 166–176, 2022.
https://doi.org/10.1007/978-3-031-11203-4_18

1 Introduction

Determining the location of anatomical structures in a mouse brain is an essential step for analyzing and understanding the architecture and function of brain circuits, and of the overall whole-brain activity [4]. Structures of interest can be located using standardized anatomical reference atlases, usually taking a two-step approach:

1. **Identification:** The input brain slice has to be identified, i.e., the corresponding 2D atlas plate has to be found.
2. **Registration:** The identified slice is registered to the corresponding atlas plate. Anatomical structures are determined based on the registered annotated plate.

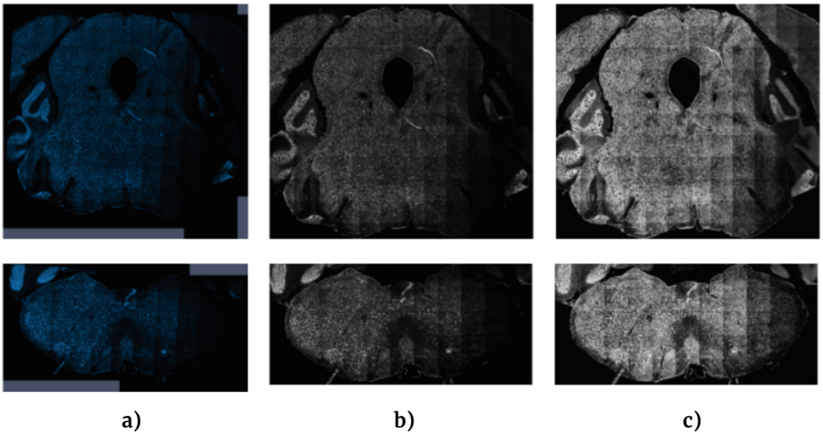


Fig. 1. a) Two typical high-resolution microscopy images showing the cross-sectional view of a mouse spinal cord in pseudo-color. The size of the input images in this work varied between 17408 × 10240 and 25600 × 20480 pixels. Notice the artefacts due to low contrast, tiling and missing regions, which make them challenging to process. b) Input images after gray scale conversion c) Pre-processed images with histogram equalization

In most cases, the acquired microscopy images of brain slices often suffer from artefacts due to missing tissue regions, irregular staining, titling errors, air bubbles and tissue wrinkles [15], as shown in Fig. 1. This is further aggravated due to additional variations in the images depending on the experimental procedures, instrumentation noise, etc. This makes it difficult to identify and register mouse brain images. For these reasons, practitioners usually resort to manually comparing image slices to 2D atlas plates which can be very time-consuming.

Compared to the registration of mouse brain images, the first part of identification has received far less attention from the brain imaging community. At the outset, wrong identification of brain slices could lead to incorrect determination of anatomical structures regardless of how well the image registration itself is

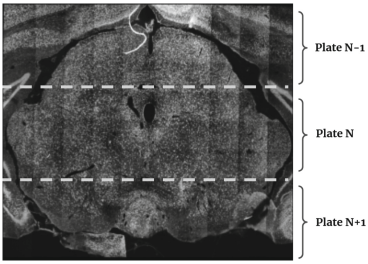


Fig. 2. Due to the difference in cutting angles compared to the atlas plates, no single ground truth plate can be registered to the input images. In this illustration, we point this out where the expert user usually would register different, usually consecutive, plates to different regions of the image.

performed. Therefore, precise determination of anatomical structures requires accurate identification of brain slices as a precursor.

The correspondence between brain slices and atlas plates could be found by reconstructing a 3D volume from the brain slices and then registering them to the 3D reference atlas [12]. However, it is not always possible to construct an accurate brain volume, e.g. when brain slices are cut at different angles or when only few brain slices are available or partial brain images are used. The difference in slice cutting angles between the atlas plates and the acquired images is a common challenge affecting the usefulness of atlas-based registration. In Fig. 2 we illustrate an instance where different regions of the same image could correspond to different atlas plates due to a mismatch between the cutting angle of the acquired image from a brain slice and the atlas plates. This way, the central region of an image corresponds to an atlas plate (Plate N) while the upper part of the image belongs to the previous plate and the bottom part to the next atlas plate. Another approach could be based on content-based image retrieval where images are queried based on some underlying image or sub-image feature descriptions [11,13].

In this study, we investigate the problem of identifying the atlas plates corresponding to mouse brain slices, when the image data are partial and/or acquired at different cutting angles. The brain slices are identified by finding the corresponding 2D coronal plates in the Allen Mouse Brain Atlas [9]. The proposed approach has similarities to some of the ideas explored within the domain of content-based image retrieval [1,13]. The brain slice identification is achieved by using convolutional neural networks (CNNs), used in the Siamese Network configuration [2,8], to obtain low-dimensional representations of the image data. These low-dimensional embeddings are contrasted with positive and negative pairs to learn a semantically meaningful space where the correspondence between brain slices and atlas plates can be determined. The image identification method is compared to SimpleElastix, which is based on the widely used tool Elastix [10], in terms of accuracy and speed.

2 Methods

Siamese Networks: In this work, CNNs are used to identify brain slices by matching them to their corresponding atlas plates. The network architecture is comprised of identical CNNs in the Siamese Network configuration [8], as shown in Fig. 3. The CNN, $S_\theta(\cdot)$, takes an image I (of height H and width W) as input and outputs a low-dimensional feature vector (embedding), h, i.e., $S_\theta(\cdot) : I \in \mathbb{R}^{H \times W} \mapsto h \in \mathbb{R}^L$, where L is the size of the embedding space and θ are the learnable network parameters. In the pairwise setting, two *sister* neural networks with shared parameters are used (Fig. 3-b).

The embeddings for brain slices, treated as the fixed image, are obtained as $h_F = S_\theta(I_F) \in \mathbb{R}^L$. The embeddings for the atlas plates, treated as the moving image, are obtained in a similar manner, $h_M = S_\theta(I_M) \in \mathbb{R}^L$. After obtaining the embeddings of the fixed and moving images, their similarity is determined based on the Euclidean distance between these embeddings, $d(h_M, h_F)$. The reference atlas plate with the lowest distance is then predicted to be the corresponding atlas plate for a given brain slice.

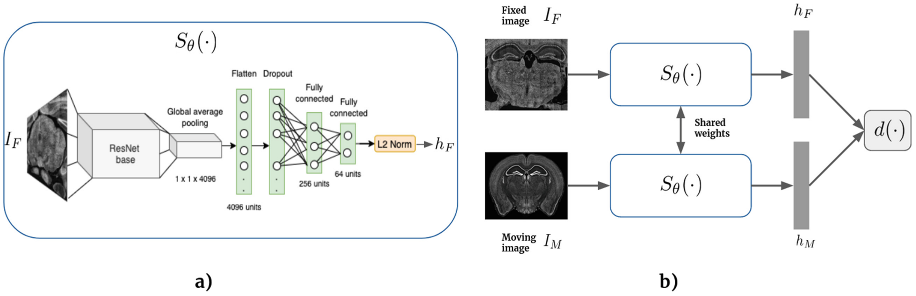


Fig. 3. a) Network architecture of the model used comprising of a ResNet-backbone and a multi-layered perceptron, $S_\theta(\cdot)$, used to obtain low-dimensional embeddings of the brain slices and atlas plates. b) Computing the similarity between brain slices and atlas plates with CNNs based on the low-dimensional representations corresponding to the moving and fixed images obtained from the identical CNNs, in a Siamese network layout, which are further used to compute their pairwise similarity, $d(\cdot)$.

Metric Learning: The distance between the embeddings of more similar images should be smaller than that between dissimilar images for the low-dimensional embedding space to be meaningful. This is achieved in this work using weakly supervised metric learning [3]. The Siamese networks for brain slice identification are trained to learn the representation of images such that corresponding brain slices and atlas plates would be closer to each other in the embedding space. We compare the embedding space learned based on training the networks with two different loss functions:

1. Contrastive loss [5], given as:

$$L = \begin{cases} \frac{1}{2} d(h_F, h_M)^2, & \text{if positive pair} \\ \frac{1}{2} \max(0, m - d(h_F, h_M))^2, & \text{if negative pair} \end{cases} \tag{1}$$

   where the positive pair is comprised of the microscopy image, I_F, and the corresponding ground truth atlas plate, I_M, and the negative pair can consist of any non-ground truth atlas plate. The parameter $m \in \mathbb{R}_+$ is the margin used to control the contribution from negative pairs.
2. Triplet loss [14], given as:

$$L = \max(d(h_A, h_P) - d(h_A, h_N) + m, 0) \tag{2}$$

   where h_A, h_P, h_N are the embeddings of anchor- (I_A), positive- (I_P) and negative- (I_N) images, respectively. Note that in case of triplet loss, a third sister network with shared weights is included to obtain feature embeddings.

Two different types of triplets (I_A, I_P, I_N) are sampled to calculate the triplet loss. These triplets are defined based on the distance between the embeddings h_A, h_P, h_N of anchor I_A, positive I_P and negative I_N images:

i) Semi-hard triplets: the distance between h_A and h_P is smaller than the distance between h_A and h_N, however, the loss is still positive.
ii) Hard triplets: the distance between h_A and h_N is smaller than the distance between h_A and h_P.

When the models are either trained with contrastive- or triplet- losses, the training process enforces structure to the embedding space so that the embeddings of similar images are pulled closer, whereas embeddings of dissimilar images are pushed away from each other. At inference, new microscopy images are ideally closer to their corresponding atlas plates in the embedding space. An overview of CNNs in Siamese network configuration for atlas plate prediction with moving and fixed images is shown in Fig. 3.

3 Data and Experiments

3.1 Data

Microscopy Data: Eighty-four high-resolution microscopy images of mouse brain slices were acquired using a 10x objective in a Zeiss LSM 900 confocal microscope from four animals. The size of the images varied between 17408 × 10240 px and 25600 × 20480 px. Most of the images were partial as they were not capturing the entire brain slice. For instance, the cortex or the cerebellar cortex were captured partially or, in some images, were not captured at all as seen in first column of Fig. 4. The images of brain slices were preprocessed, cropped and equalized using Contrast Limited Adaptive Histogram Equalization

(CLAHE) to reduce some artefacts, as shown in Fig. 1. The dataset was split into four sets: training (50 images), validation-1 (12 images), validation-2 (10 images) and test (12 images).

Ground Truth: The Allen Mouse Brain Atlas [9] was used as the reference atlas. It consisted of 132 Nissl-stained coronal plates spaced at 100 µm, seen in the second column of Fig. 4. The ground truth in these experiments were the atlas plate numbers which were provided by a neuroscientist with expertise in manual registration of these images. For a given brain slice, there could be several matching plates due to the difference in cutting angles, as shown in Fig. 2. However, the domain expert marked a single plate to be the ground truth depending on whichever plate best described specific regions of interest. This is to say, in most applications involving these data there are no hard ground truths as each slice could correspond to several consecutive atlas plates due to the difference in cutting angles.

Data Augmentation: To capture variations in the microscopy data beyond the limited training set extensive data augmentation (affine transformation, cropping and padding, pepper noise) was applied to the training dataset. Data augmentation was performed on all the 50 training set brain slices and also the 132 atlas plates. In order to reduce computations, the high resolution images were resized to square inputs of size 1024^2, 512^2 or 224^2 depending on the experiment.

3.2 Experiments

Experiments: The performance of our CNN-based slice identification method was compared with a baseline SimpleElastix-based algorithm that identifies brain slices based on mutual information (MI). The baseline method affinely registers each brain slice with every atlas plate and picks the atlas plate with the highest

Table 1. Mean Absolute Error (MAE) on the *validation-2* dataset for identifying brain slices with our method. The lowest MAE is achieved by the network with ResNet50v2 base, trained with semi-hard triplet loss and using 1024^2 images. B is the training batch size.

Loss	B	ResNet50v2			ResNet101v2		
		224^2	448^2	1024^2	224^2	448^2	1024^2
Triplet (semi-hard)	32	2.5	2.2	2.8	1.9	3.1	3.1
	16	2.0	3.7	**1.8**	2.6	2.1	2.7
Triplet (hard)	32	2.4	3.0	3.0	2.8	3.7	2.7
	16	3.1	2.8	2.6	2.0	2.7	2.8
Contrastive	32	3.6	2.1	3.4	4.2	2.5	5.6

MI. In total, 100 random hyperparameters from the SimpleElastix affine parameter map were tested. The results of the best performing baseline model (with 7 resolutions using recursive image pyramid and random sample region, 2800 iterations in each resolution level and disabled automatic parameter estimation) are used for comparison.

Metrics: The methods were evaluated based on three metrics: Mean Absolute Error (MAE), TOP-N accuracy and inference time. MAE measured the accuracy of predictions. For each brain slice all 132 atlas plates were ranked (starting from zero) based on the similarity score (the Euclidean distance or MI, depending on the method). Then MAE was computed as $MAE = (\sum_{i=0}^{N} y_i)/N$, where N is the number of brain slices, y_i is the position of ranked ground truth atlas plate for a given brain slice i. With 132 atlas plates used, MAE can have values in the range $[0, 131]$. If all brain slices are identified correctly, MAE is equal to 0. To account for the inherent ambiguity in ground truth we report Top-3, Top-5 and Top-10 accuracy.

Hyperparameters: Fig. 3-a) shows the architecture of the Siamese Networks with the embedding space feature dimension $L = 64$. The base of network consists of a CNN-backbone implemented as ResNet network [6] pre-trained on the ImageNet dataset. The CNN backbone is followed by a multi-layered perceptron that outputs the embedding. While training the networks, all layers of the ResNets were *frozen* except the last ones starting with the prefix *conv5*. The networks were trained on the *training* dataset for a maximum of 10 k iterations using the Adam optimizer [7] with an initial learning rate of 10^{-4}. The experiments were performed on Nvidia GeForce RTX 3090 GPU, i7-10700F CPU and 32 GB memory. The training was stopped if MAE on the *validation-1* dataset was not decreasing for more than 2 k iterations.

Results: The converged models based on *validation-1* set were evaluated on the *validation-2* dataset, and the MAE performance for two ResNet backbones (ResNet50, ResNet101), the various loss functions, input- and batch- sizes are reported in Table 1. The best performing configuration is the ResNet50 backbone network trained with batch size (B) of 16 using input size 1024^2 with the semi-hard triplet loss with MAE=1.8. This best performing model was further evaluated on the *test* dataset and compared with the SimpleElastix-based approach, reported in Table 2. We notice that the MAE on test set for our method is 1.42 compared to 60.4 for the baseline. Our method obtained Top-3 accuracy of obtaining 83.3% compared to 25% for the baseline. A similar trend is observed for Top-5 and Top-10 accuracy, where our method achieves 100% accuracy. The total inference time on the *test* set for the two methods are also reported in Table 2 where we observe that the baseline method takes orders of magnitude more time than the trained CNN model.

Finally, the Top-5 predicted atlas plates on a subset of the *test* dataset are reported in Table 3. In all the cases, the ground truth plate is within the Top-5 predictions highlighted in bold. Examples of the predicted atlas plates by our method that have the highest similarity are visualized in Fig. 4.

Table 2. Performance of our CNN-based method compared to the SimpleElastix-based approach on the *test* dataset for identifying brain slices reported as Top-N accuracy. Our method trained with semi-hard triplet loss outperforms SimpleElastix-based approach by a large margin in all the evaluated metrics. Inference time measures the time taken to identify all 12 brain slices from the *test* dataset.

	MAE	TOP-1	TOP-3	TOP-5	TOP-10	Infer. time
SimpleElastix	60.4	16.7%	25%	25%	25%	12h 25 m
Siamese Networks	**1.42**	**25%**	**83.3%**	**100%**	**100%**	**7.2 s**

Table 3. Identifying brain slices from the subset of the *test* dataset: the labels of ground truth and Top-5 predicted atlas plates by our CNN-based method. Even though some predictions are incorrect, all of them are close to the ground truth labels. Labels define the position of atlas plates in the reference atlas.

Ground truth	Top-5 predictions
91	92, **91**, 93, 90, 94
130	129, 128, **130**, 131, 127
86	87, 88, **86**, 85, 89
63	62, 61, 60, **63**, 59
108	109, 110, 111, 112, **108**

4 Discussions and Conclusions

Our CNN-based method in the Siamese network configuration used to identify brain slices have shown impressive results, i.e. in finding corresponding coronal 2D atlas plates. Our method performed well even when most images were missing image regions, and some images belonging to different classes (plate numbers) looked very similar to each other, thus making the identification task even more complex. Training with contrastive- and triplet- losses solve this issue by using margin, i.e., dissimilar images are not pushed away if the distance between them is larger than the margin.

The identification accuracy (MAE) had no clear correlation with the batch size (16 and 32), the image resolution (224 × 224, 448 × 448, 1024 × 1024) and the type of the base for the Siamese network (ResNet50v2 and ResNet101v2), as seen in Table 1. However, using images with lower resolution and networks with fewer parameters could further improve the inference time. We did not observe the performance of our method to be highly influenced by the choice of loss functions. The models trained with triplet loss rather than contrastive loss, on average, achieved higher accuracy, however, the difference is not significant.

Evaluating the performance of the method using ambiguous ground truth data due to variations in cutting angle was another challenge. This was overcome by evaluating the methods using Top-N accuracy instead of only predicting the most similar atlas plate. We observe that our method achieved TOP-5 accuracy

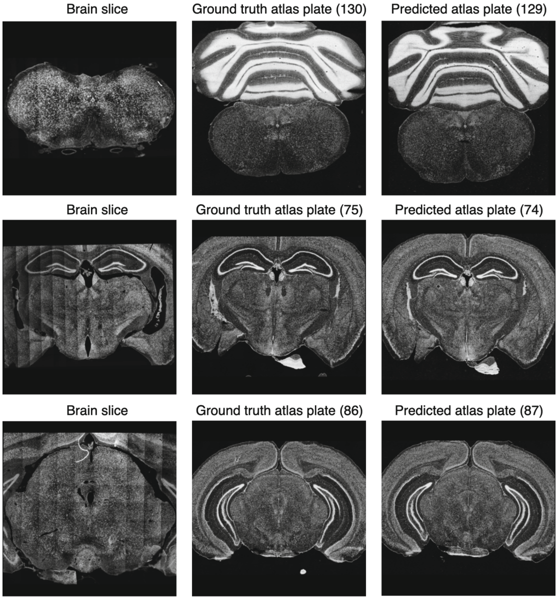

Brain slice
Ground truth atlas plate (130)
Predicted atlas plate (129)
Brain slice
Ground truth atlas plate (75)
Predicted atlas plate (74)
Brain slice
Ground truth atlas plate (86)
Predicted atlas plate (87)

Fig. 4. Examples of the predicted (most similar) atlas plates by our method. Note that in all cases the ground truth plates are predicted within the Top-5 candidates in Table 3. Columns: **(1)** brain slices from the test dataset; **(2)** ground truth atlas plates; **(3)** predicted atlas plates. The number in parentheses shows the label of the atlas plate, i.e. the position of the atlas plate in the reference atlas.

of 100% meaning that the actual corresponding atlas plate always falls in the top 5 predicted atlas plates, as seen in Table 2. Further, the variations within the Top-5 predictions for all five cases reported in Table 3 could be plausible, as most of the predictions are neighbouring atlas plates of the ground truth. We also report the Top-1 accuracy and notice a drop in performance for both methods due to the inherent ambiguity in the ground truth. The inherent ambiguity of the ground truth makes our method more useful as practitioners can explore several likely candidate atlas plates to register to.

In conclusion, we proposed to use CNNs in Siamese Network configuration trained with contrastive- and triplet- losses as a method for identifying correspondence between complete and partial mice brain slices. Challenges such as partial/missing data and variations in cutting angles were overcome by learning a semantically meaningful embedding space. Our method has shown large performance improvements in both accuracy and inference times compared to the SimpleElastix-based baseline. With this work, we have we demonstrated the usefulness of this approach with a 2D reference atlas. We hypothesize that the same method can also be applied to a 3D reference atlas for further improved precision in the slice identification task.

5 Discussions and Conclusions

The Siamese networks used to identify brain slices has shown impressive results, i.e. in finding corresponding coronal 2D atlas plates. It achieved TOP-5 accuracy of 100% meaning that the actual corresponding atlas plate always falls in the top 5 predicted atlas plates. The identification accuracy (MAE) had no clear correlation with the batch size (16 and 32), the image resolution (224 × 224, 448 × 448, 1024 × 1024) and the type of the base for the Siamese network (ResNet50v2 and ResNet101v2). However, using images with lower resolution and networks with fewer parameters could improve the inference time. We did not observe that the performance of the Siamese network would be highly influenced by the loss function, namely contrastive and triplet losses. The models trained with triplet loss rather than contrastive loss, on average, achieved higher accuracy, however, the difference is not significant.

The Siamese networks produced impressive results even though some images of different classes looked very similar to each other, thus making the identification task even more complex. The distance between such images should be lower than the distance between two completely dissimilar images. Maximizing the distance between all images of different classes would make it difficult for networks to learn representations of these classes. Contrastive and triplet losses solve this issue by using margin, i.e., dissimilar images are not pushed away if the distance between them is larger than the margin.

In this study, we proposed Siamese Networks as a method for identifying complete and partial mouse brain slices, i.e. finding the corresponding 2D atlas plates. The networks have shown a high precision and significantly improved inference time compared to the baseline. While we demonstrated this with a 2D reference atlas, the same method can also be applied to a 3D reference atlas for even higher identification precision.

6 Conclusions

In this study, we proposed Siamese Networks as a method for identifying complete and partial mouse brain slices, i.e. finding the corresponding 2D atlas plates. The networks have shown a high precision and significantly improved inference time compared to the baseline. While we demonstrated this with a 2D reference atlas, the same method can also be applied to a 3D reference atlas for even higher identification precision.

Acknowledgments. The authors thank Kiehn Lab (University of Copenhagen, Denmark) for providing access to the microscopy images and the hardware used to train the models. They also acknowledge the Core Facility for Integrated Microscopy (CFIM) at the Faculty of Health and Medical Sciences for support with image acquisition.

Compliance with Ethical Standards. All animal experiments and procedures were carried according to the EU Directive 2010/63/EU and approved by the Danish Animal Experiments Inspectorate (Dyreforsøgstilsynet) and the Local Ethics Committee at the University of Copenhagen.

References

1. Breznik, E., Wetzer, E., Lindblad, J., Sladoje, N.: Cross-modality sub-image retrieval using contrastive multimodal image representations. arXiv preprint arXiv:2201.03597 (2022)
2. Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R.: Signature verification using a Siamese time delay neural network. Adv. Neural Inf. Process. Syst. **6**, 737–744 (1993)
3. Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), vol. 1, pp. 539–546. IEEE (2005)
4. Furth, D., et al.: An interactive framework for whole-brain maps at cellular resolution. Nat. Neurosci. **21**(1), 139–149 (2017). https://doi.org/10.1038/s41593-017-0027-7
5. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), vol. 2, pp. 1735–1742 (2006). https://doi.org/10.1109/CVPR.2006.100
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
7. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization (2017)
8. Koch, G., Zemel, R., Salakhutdinov, R.: Siamese neural networks for one-shot image recognition (2015)
9. Lein, E.S., Hawrylycz, M.J., Ao, N.: Genome-wide atlas of gene expression in the adult mouse brain. Nature **445**(7124), 168–176 (2007). https://doi.org/10.1038/nature05453, https://www.nature.com/articles/nature05453
10. Marstal, K., Berendsen, F., Staring, M., Klein, S.: SimpleElastix: a user-friendly, multi-lingual library for medical image registration. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 574–582 (2016). https://doi.org/10.1109/CVPRW.2016.78
11. Müller, H., Michoux, N., Bandon, D., Geissbuhler, A.: A review of content-based image retrieval systems in medical applications-clinical benefits and future directions. Int. J. Med. Inform. **73**(1), 1–23 (2004)
12. Pichat, J., Iglesias, J.E., Yousry, T., Ourselin, S., Modat, M.: A survey of methods for 3D histology reconstruction. Med. Image Anal. **46**, 73–105 (2018). https://doi.org/10.1016/j.media.2018.02.004
13. Qayyum, A., Anwar, S.M., Awais, M., Majid, M.: Medical image retrieval using deep convolutional neural network. Neurocomputing **266**, 8–20 (2017)
14. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: a unified embedding for face recognition and clustering. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015. https://doi.org/10.1109/cvpr.2015.7298682, http://dx.doi.org/10.1109/CVPR.2015.7298682
15. Xiong, J., Ren, J., Luo, L., Horowitz, M.: Mapping histological slice sequences to the Allen mouse brain atlas without 3D reconstruction. Front. Neuroinform. **12**, 93 (2018). https://doi.org/10.3389/fninf.2018.00093, https://www.frontiersin.org/article/10.3389/fninf.2018.00093

Transformed Grid Distance Loss for Supervised Image Registration

Xinrui Song[1], Hanqing Chao[1], Sheng Xu[2], Baris Turkbey[3],
Bradford J. Wood[2], Ge Wang[1], and Pingkun Yan[1(✉)]

[1] Department of Biomedical Engineering and Center for Biotechnology
and Interdisciplinary Studies, Rensselaer Polytechnic Institute, Troy, NY 12180, USA
yanp2@rpi.edu

[2] Center for Interventional Oncology, Radiology and Imaging Sciences,
National Institutes of Health, Bethesda, MD 20892, USA

[3] Molecular Imaging Program, National Cancer Institute, National Institutes
of Health, Bethesda, MD 20892, USA

Abstract. Many deep learning image registration tasks, such as volume-to-volume registration, frame-to-volume registration, and frame-to-volume reconstruction, rely on six transformation parameters or quaternions to supervise the learning-based methods. However, these parameters can be very abstract for neural networks to comprehend. During the optimization process, ill-considered representations of rotation may even trap the objective function at local minima. This paper aims to expose these issues and propose the Transformed Grid Distance loss as a solution. The proposed method not only solves the problem of rotation representation but unites the gap between translation and rotation. We test our methods both with synthetic and clinically relevant medical image datasets. We demonstrate superior performance in comparison with conventional losses while requiring no alteration to the network input, output, or network structure at all.

1 Introduction

Existing deep learning-based image registration methods have explored many types of supervision. Unsupervised methods such as [1,4,11] relies on image intensity-based similarity metrics to supervise the network. These methods, however, are limited to single-modality registration tasks, or multi-modal images with very similar content and texture. Weakly supervised registration [2,7] incorporated weak labels such as organ segmentation to guide the training process.

In contrast, supervised methods require the ground truth annotations of registration for training [3,5]. For deformable image registration, providing such annotations can be unrealistically difficult. However, for tasks in which no significant differences were found between rigid and deformable registrations [10], using rigid registration reduces the annotation cost significantly. For example, in image-fusion guided prostate cancer biopsies, the manual registration between the MR and ultrasound images has been a routine for the clinical procedure.

A. Hering et al. (Eds.): WBIR 2022, LNCS 13386, pp. 177–181, 2022.
https://doi.org/10.1007/978-3-031-11203-4_19

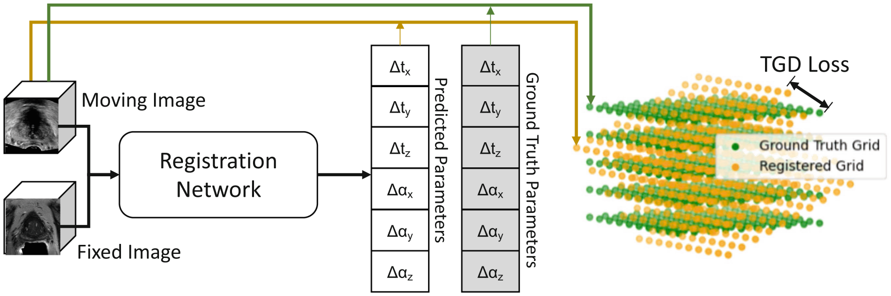


Fig. 1. Illustration of the transformed grid distance (TGD) loss.

Requesting these manual registration labels for training come at no additional cost to the clinicians. In other scenarios where deformable registration is preferred, a rigid transformation is also often required to pre-align the images before any deformable registration can be performed. For the above reasons, supervised deep learning based rigid image registration has been intensively studied, and will be the focus of this work.

Labels and loss function are critical components of supervised image registration. Since 3D rigid transformation is commonly represented by six transformation parameters, including three rotation angles and a 3D translation vector, a straightforward option is to use the distance between the ground truth and estimated transformation parameters as the loss to train the image registration network. However, numerous works [6,8,12] pointed out that the Euler angle representation is problematic for loss computation. In some cases, quaternion angles are used instead. In this paper, we argue that neither of them is the optimal choice for being used directly in a loss function. Instead, these abstract mathematical expressions should be first converted into more physically intuitive values. We propose a new loss – the Transformed Grid Distance (TGD) loss for network training.

2 Transformed Grid Distance

In supervised rigid registration, transformation parameters are often used as the label for network supervision. Compact transformation parameters, either in Euler or quaternion representation, can be difficult for neural networks to learn through conventional loss functions (e.g. L1 and MSE loss).

Instead of directly supervising the transformation parameters themselves, we apply the estimated transformation on the moving image grid, and supervise the distance between the transformed points and their corresponding points in the ground truth grid as illustrated in Fig. 1. Let $\boldsymbol{G} \in \mathbb{R}^{m \times n \times l}$ denote a 3D moving image grid. TGD loss is computed as

$$L_{TGD} = \left\| T_\theta(\boldsymbol{G}) - T_{gt}(\boldsymbol{G}) \right\|_2 , \tag{1}$$

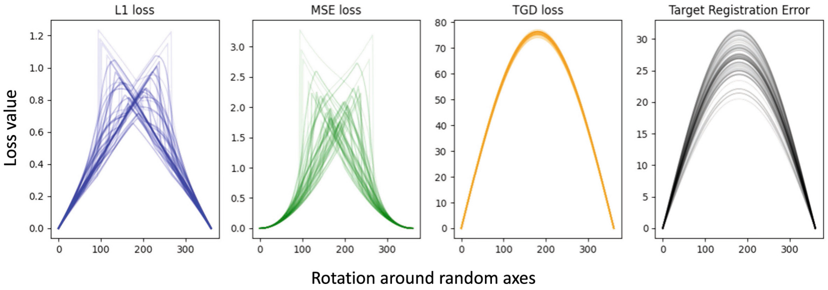


Fig. 2. We generated the rotations in this figure by randomly selecting 27 unit vectors and ranging the rotation amplitude from 0 to 360 °C. The Target Registration Error serves as the evaluation metric, as in many registration tasks.

where T denotes a 3D transformation matrix converted from θ. The key difference here is to convert the abstract representation θ into a dense and intuitive representation, which guides the network optimization process through circumventing any non-linear transformation conversions that the network would otherwise have to figure out.

The proposed TGD loss elegantly unifies both rotation and transformation into point-wise distance, which results in a smooth loss landscape that guides the network learning process. Had more meaningful points been acquired (*i.e.* anatomical landmarks), the loss can be simply adapted into Target Registration Error (TRE) by replacing the grid with those clinically relevant points. One major weakness of the Euler angles is that they must be applied in a fixed order, which is not reflected at all with L1 or MSE loss. During training, each line from Fig. 2 can be regarded as a training sample. The loss curves for either L1 or MSE loss on Euler angles vary wildly from sample to sample, while the proposed TGD loss stays consistent with the Target Registration Error (TRE).

The quaternion system seems to be a better solution than the Euler angles. However, due to the fact that the quaternion expression is divided into two intertwined parts, it is hard to guarantee that the direction of optimization is at all correct. For example, slight error in the rotation axis would result in a large TRE regardless of the rotation angle.

3 Experiments

In this section, we present both a synthetic and a clinically relevant experiment. Our dataset consists of 528 manually labeled cases of MR-transrectal ultrasound (TRUS) volume pair for training, 66 cases for validation, and 68 cases for testing.

In the first experiment, we use an MR volume as the fixed image, and its own perturbed result as the moving image. We have also included the result of TRE-TGD loss, which is another version of the proposed method that replaces the regular grid points in TGD loss with the target prostate surface points. The quaternion loss, on the other hand, failed to converge in this experiment where

Table 1. Performance of different loss functions in MR-MR registration.

Method	Mean TRE (mm)	Percentiles [25th, 50th, 75th, 95th]
Initial	12.66 ± 7.30	[6.39, 12.83, 18.79, 23.88]
Quaternion loss	12.95 ± 7.36	[6.64, 12.93, 19.00, 24.67]
MSE Euler angle loss	2.68 ± 2.31	[1.19, 2.04, 3.43, 6.82]
L1 Euler angle loss	2.80 ± 2.68	[1.04, 2.09, 3.72, 7.48]
TGD loss	$\mathbf{1.51 \pm 1.45}$	[0.62, 1.15, 1.93, 3.85]
TRE-TGD loss	$\mathbf{1.50 \pm 1.53}$	[0.65, 1.14, 1.87, 3.83]

Table 2. Performance of different loss functions in MR-TRUS registration.

Method	Mean TRE (mm)	Percentiles [25th, 50th, 75th, 90th]
Initial	9.93 ± 5.87	[4.89, 9.82, 14.89, 19.10]
MSE Euler angle loss	5.57 ± 2.86	[3.47, 5.06, 7.07, 10.98]
SRE-TGD loss	$\mathbf{4.40 \pm 2.49}$	[2.57, 3.97, 5.77, 8.88]

large rotation errors are concerned. Results in Table 1 show that simply through 'rephrasing' the transformation parameters into physical distance between grid points, the network was guided to converge at a lower minimum.

The second experiment treats the TRUS volume as the moving image, and the corresponding MR volume as the fixed image. This is a use case, where an accurate alignment between the transrectal ultrasound (TRUS) and MR volume greatly benefits the ultrasound-guided prostate cancer biopsy [9]. For each pair of MR and TRUS volume, we are provided with the manual label for rigid registration from TRUS to MR, as well as the prostate surface points in MR. Similar to the TRE-TGD loss in the first experiment, the SRE-TRD loss also calculates the distance between corresponding points, thereby a subset to the proposed TGD loss. Table 2 compares the result of multi-modal registration between the conventional MSE loss and SRE-TGD loss. With the same network architecture and other settings, the proposed loss function results in a significant ($p < 0.001$ under paired t-test) improvement over the conventional MSE loss.

4 Discussions and Conclusion

In this paper, we revealed the limitation of directly using abstract transformation parameters for loss computation in supervised training of image registration networks. With such insight, we introduced a simple yet effective tool to boost the performance of supervised rigid volume registration. Although the analysis and experiments are mainly conducted in a rigid setting, this idea can be easily adapted for a non-rigid affine registration task.

References

1. 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)
2. Baum, Z.M.C., Hu, Y., Barratt, D.C.: Multimodality biomedical image registration using free point transformer networks. In: Hu, Y., et al. (eds.) ASMUS/PIPPI - 2020. LNCS, vol. 12437, pp. 116–125. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-60334-2_12
3. Guo, H., Kruger, M., Xu, S., Wood, B.J., Yan, P.: Deep adaptive registration of multi-modal prostate images. Comput. Med. Imaging Graph. **84**, 101769 (2020)
4. Hansen, L., Heinrich, M.P.: GraphRegNet: deep graph regularisation networks on sparse keypoints for dense registration of 3D lung CTs. IEEE Trans. Med. Imaging **40**(9), 2246–2257 (2021)
5. Haskins, G., et al.: Learning deep similarity metric for 3D MR-TRUS image registration. Int. J. Comput. Assist. Radiol. Surg. **14**(3), 417–425 (2019)
6. Hou, B., et al.: 3-D reconstruction in canonical co-ordinate space from arbitrarily oriented 2-D images. IEEE Trans. Med. Imaging **37**(8), 1737–1750 (2018)
7. Hu, Y., et al.: Weakly-supervised convolutional neural networks for multimodal image registration. Med. Image Anal. **49**, 1–13 (2018)
8. Kendall, A., Grimes, M., Cipolla, R.: PoseNet: a convolutional network for real-time 6-DOF camera relocalization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2938–2946 (2015)
9. Song, X., et al.: Cross-modal attention for MRI and ultrasound volume registration. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12904, pp. 66–75. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87202-1_7
10. Venderink, W., de Rooij, M., Sedelaar, J.M., Huisman, H.J., Fütterer, J.J.: Elastic versus rigid image registration in magnetic resonance imaging-transrectal ultrasound fusion prostate biopsy: a systematic review and meta-analysis. Eur. Urol. Focus **4**(2), 219–227 (2018)
11. de Vos, B.D., Berendsen, F.F., Viergever, M.A., Staring, M., Išgum, I.: End-to-End unsupervised deformable image registration with a convolutional neural network. In: Cardoso, M.J., et al. (eds.) DLMIA/ML-CDS -2017. LNCS, vol. 10553, pp. 204–212. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67558-9_24
12. Wei, W., Haishan, X., Alpers, J., Rak, M., Hansen, C.: A deep learning approach for 2D ultrasound and 3D CT/MR image registration in liver tumor ablation. Comput. Methods Programs Biomed. **206**, 106117 (2021)

Efficiency

Deformable Lung CT Registration by Decomposing Large Deformation

Jing Zou[1], Lihao Liu[2], Youyi Song[1], Kup-Sze Choi[1], and Jing Qin[1](✉)

[1] Centre for Smart Health, School of Nursing, The Hong Kong Polytechnic University, Hong Kong, China
harry.qin@polyu.edu.hk

[2] Centre for Mathematical Sciences, University of Cambridge, Cambridge, UK

Abstract. Deformable lung CT registration plays an important role in image-guided navigation systems, especially in the situation with organ motion. Recent progress has been made in image registration by utilizing neural networks for end-to-end inference of a deformation field. However, there are still difficulties to learn the irregular and large deformation caused by organ motion. In this paper, we propose a patient-specific lung CT image registration method. We first decompose the large deformation between the source image and the target image into several continuous intermediate fields. Then we compose these fields to form a spatio-temporal motion field and refine it through an attention layer by aggregating information along motion trajectories. The proposed method can utilize the temporal information in a respiratory circle and can generate intermediate images which are helpful in image-guided systems for tumor tracking. Extensive experiments were performed on a public dataset, showing the validity of the proposed methods.

Keywords: Image registration · Lung CT · Organ movement · Deformation field decomposition · Attention layer

1 Introduction

Image-guided navigation systems have greatly enhanced the therapeutic efficiency of complicated interventions [1]. However, in such systems, organ motions caused by respiration is a major challenge of accurate lesion targeting. In current practice, this challenge is often handled by asking the patients to hold their breath and scanning repeated CTs. This will either cause distress to patients or increase the radiation exposure. To the end, registration is a promising technique to correct the position offset of the targeting organ or tumor.

Recently, deep networks have been applied to address deformable registration problems and achieved remarkable success [3,6–10]. However, it is still difficult to accurately estimate the large deformation due to respiration (tumors and sensitive structures in the thorax can move more than 20 mm [12]). In this paper, we propose a lung CT registration method that utilizes temporal information

A. Hering et al. (Eds.): WBIR 2022, LNCS 13386, pp. 185–189, 2022.
https://doi.org/10.1007/978-3-031-11203-4_20

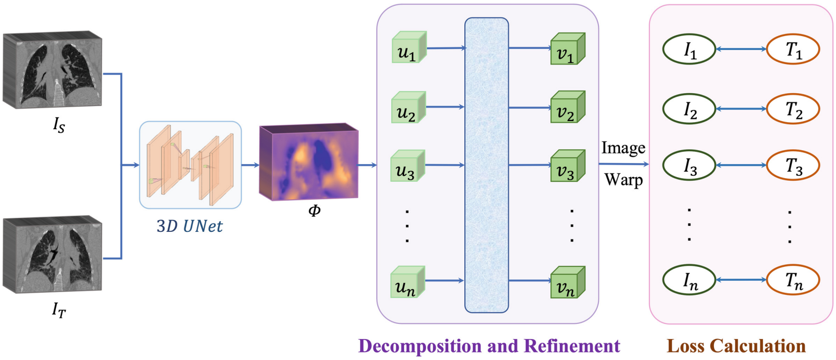


Fig. 1. The illustrative pipeline of our method.

during respiration. During training, the images at extreme phases , as well as the intermediate images, are employed as training data. Once the network is trained, it can infer a deformation field without the intermediate images.

2 Methodology

Let I_S and $I_T \in \mathbb{R}^{H \times W \times C}$ be the source and target lung CT images, respectively. Our aim is to figure out the deformation filed $\varPhi \in \mathbb{R}^{H \times W \times C \times 3}$ that stores the coordinate offset between I_S and I_T. We employ a deep network f that takes I_S and I_T as the input to predict $\varPhi$ by solving the below problem:

$$\underset{f \in \mathcal{F}}{\operatorname{argmin}}\, \ell(I_T, I_S \circ \varPhi_f) + \lambda \mathcal{R}(\varPhi_f), \tag{1}$$

where $\mathcal{F}$ denotes the function space of f and $\varPhi_f$ stands for $\varPhi$ with f given the input (I_S, I_T). $I_S \circ \varPhi_f$ represents I_S warped by $\varPhi_f$, and ℓ is the loss function to measure the discrepancy between I_T and $I_S \circ \varPhi_f$. $\mathcal{R}(\varPhi_f)$ stands for the regularization term with the hyperparameter λ to balance its importance.

This training paradigm can work well when the deformation of lungs is small [4], but fail for a large and irregular deformation, in which pixels are dramatically deformed, diminishing the accuracy of the registration. Our method aims to solve this issue by decomposing the deformation field into several ones with small deformations and gradually refining them through an attention layer. An overview of the proposed method is shown in Fig. 1.

Decomposition. We first decompose the deformation field $\varPhi$. This field describes the directions and the distances for all voxels moving from I_S to I_T. Considering the progressive movement of the lung, the deformation field can be decomposed by incremental steps to obtain intermediate deformation fields u_i.

We assume that each voxel deforms along a straight line [11]. Thus the decomposition can be achieved by linear interpolation: $u_i = \Phi/n$, where n denotes the phases in a respiratory circle.

Refinement. Above mentioned linear interpolation of the deformation field relies on the assumption that the displacement of each voxel is homogeneous. However, in practice, the deformation may be irregular. So we refine these small deformation fields by firstly concatenating them to form a spatio-temporal motion field U, which contains spatial and temporal information during respiration. Then we input the motion field U to a self-attention layer, and output the refined field V. At last, V is decomposed again to obtain refined intermediate fields v_i, with which I_S are warped to generate intermediate images I_n that are used to calculate loss with ground truth intermediate images T_n. Finally, our decomposition method aims to train the deep network f for deformable registration by solving the following problem:

$$\operatorname*{argmin}_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell_t(T_t, I_S \circ (t\Phi_f/n)) + \lambda \mathcal{R}(\Phi_f). \tag{2}$$

3 Experiments

Experimental Setup: Our method was evaluated on a public dataset [5], which has ten thoracic 4D CTs obtained at ten different respiratory phases in a respiratory cycle. In each 4D CT, 300 anatomical landmarks were manually annotated at two extreme phases. We evaluate our method with target registration error (TRE), which is formulated as the average Euclidean distance between the fixed landmarks and the warped moving landmarks. We implemented our method with Pytorch on an NVIDIA RTX 3090 GPU.

Experimental Results: We compare our method with five competitive methods: BL [2] (CVPR 2018), IL [6] (MedIA 2019), VM [3] (TMI 2019), MAC [7] (MedIA 2021), and CM [8] (MedIA 2021), denoting the baseline and existing methods that use iterative learning strategy, lung masks as the supervision, landmarks as the supervision, and the cycle consistency, respectively. For a fair comparison, we employed the same backbone network (3D UNet) and the same learning setting.

The results via cross-validation are reported in Table 1. We can see that our method achieved the best performance of the average *TRE* (denoted as *Ave.* in the table). It improves the performance over the second-best method (VM) with 8.0%. This demonstrates the validity of our method. We also can see that our method works consistently well in the best and worst cases (denoted as *Best* and *Worst*). Moreover, the performance of our method is less diverse than others as we have the lowest *Std.* (1.06). These evidences suggest that our algorithm is more reliable and effective. We also checked the statistical significance of the performance improvement by paired t-test. We can see that, expect VM (whose p-value is 0.053), other p-values are less than 0.05, which implies that our method significantly improves the registration performance.

Table 1. The *TRE* (*mm*) results of our algorithm and compared methods.

	BL	IL	VM	MAC	CM	Ours
Ave.	3.53	3.85	3.38	3.53	3.56	**3.11**
Std.	1.38	1.25	1.17	1.25	1.56	**1.06**
Best	1.97	2.19	2.19	2.19	1.97	**1.75**
Worst	6.02	5.93	5.79	6.27	6.77	**4.8**
p-value	0.015	0.001	0.053	0.022	0.045	—

4 Conclusion

In this paper, we have investigated a simple and effective method to learn the large deformation field in lung CT image registration, which is helpful in image-guided navigation systems. This method decomposes the large deformation field into small fields, and then composes these small fields and refines them by attention layer. The experimental results show that our method works better than existing methods.

Acknowledgement. The work described in this paper is supported by a grant of Hong Kong Research Grants Council under General Research Fund (no. 15205919).

References

1. Anzidei, M., et al.: Preliminary clinical experience with a dedicated interventional robotic system for CT-guided biopsies of lung lesions: a comparison with the conventional manual technique. Eur. Radiol. **25**(5), 1310–1316 (2015)
2. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: An unsupervised learning model for deformable medical image registration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9252–9260 (2018)
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. Beg, M.F., Miller, M.I., Trouvé, A., Younes, L.: Computing large deformation metric mappings via geodesic flows of diffeomorphisms. Int. J. Comput. Vis. **61**(2), 139–157 (2005)
5. Castillo, R., et al.: A framework for evaluation of deformable image registration spatial accuracy using large landmark point sets. Phys. Med. Biol. **54**(7), 1849 (2009)
6. 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. Med. Image Anal. **52**, 128–143 (2019)
7. Hering, A., Häger, S., Moltz, J., Lessmann, N., Heldmann, S., van Ginneken, B.: CNN-based lung CT registration with multiple anatomical constraints. Med. Image Anal. **72**, 102139 (2021)

8. Kim, B., Kim, D.H., Park, S.H., Kim, J., Lee, J.G., Ye, J.C.: CycleMorph: cycle consistent unsupervised deformable image registration. Med. Image Anal. **71**, 102036 (2021)
9. Liu, L., Aviles-Rivero, A.I., Schönlieb, C.B.: Contrastive registration for unsupervised medical image segmentation. arXiv preprint arXiv:2011.08894 (2020)
10. Liu, L., Huang, Z., Liò, P., Schönlieb, C.B., Aviles-Rivero, A.I.: Pc-SwinMorph: patch representation for unsupervised medical image registration and segmentation. arXiv preprint arXiv:2203.05684 (2022)
11. Sarrut, D., Boldea, V., Miguet, S., Ginestet, C.: Simulation of four-dimensional CT images from deformable registration between inhale and exhale breath-hold CT scans. Med. Phys. **33**(3), 605–617 (2006)
12. Schreibmann, E., Chen, G.T., Xing, L.: Image interpolation in 4D CT using a BSpline deformable registration model. Int. J. Radiat. Oncol. Biol. Phys. **64**(5), 1537–1550 (2006)

You only Look at Patches: A Patch-wise Framework for 3D Unsupervised Medical Image Registration

Lihao Liu(✉), Zhening Huang, Pietro Liò, Carola-Bibiane Schönlieb, and Angelica I. Aviles-Rivero

University of Cambridge, Cambridge, UK
ll610@cam.ac.uk

Abstract. Medical image registration is a fundamental task for a wide range of clinical procedures. Automatic systems have been developed for image registration, where the majority of solutions are supervised techniques. However, those techniques rely on a large and well-representative corpus of ground truth, which is a strong assumption in the medical domain. To address this challenge, we propose a novel unified unsupervised framework for image registration and segmentation. The highlight of our framework is that patch-based representation is key for performance gain. We first propose a patch-based contrastive strategy that enforces locality conditions and richer feature representation. Secondly, we propose a patch stitching strategy to eliminate artifacts. We demonstrate, through our experiments, that our technique outperforms current state-of-the-art unsupervised techniques.

1 Introduction

Image registration seeks to find a mapping that aligns an unaligned image to a reference one. The estimated spatial mapping aims to best align the anatomical structure of interest. Majority of existing works have been investigated from the classic perspective. Whilst promising performance has been reported, those techniques build upon costly optimisation schemes, which limits their efficiency when using a large volume of data. This limitation has encouraged the fast development of deep learning techniques for medical image registration. A set of techniques have been reported based on supervised learning. However, the need for a well-representative and high-quality ground truth is a strong assumption and hard to obtain in the medical domain. Another set of techniques have been devoted to explore unsupervised techniques e.g. [1–4,7,8]. Existing techniques have proposed several network mechanisms and explicit regularisers, to accommodate a certain level, with the lack of prior knowledge. However, the performance is still limited due to the lack of high-quality prior knowledge.

Our work is motivated by the aforementioned limitation. We argue that better quality prior can be estimated from patches rather than the full image. Medical images have complex anatomical structures, which impose a challenge when

A. Hering et al. (Eds.): WBIR 2022, LNCS 13386, pp. 190–193, 2022.
https://doi.org/10.1007/978-3-031-11203-4_21

estimating an image-to-image mapping. Therefore, our modeling hypothesis is that patch embeddings are more meaningful representation for performance gain. In this work, we introduce a novel unified framework for unsupervised image registration and segmentation, which we call PC-SwinMorph (**P**atch **C**ontrastive Strategy with **S**hifted-**win**dow multi-head self-attention based on Voxel**Morph**). We underline two major highlights of our framework. Firstly, we introduce a patchwise contrastive registration strategy for richer feature representation. Secondly, we propose a patch stitching strategy to address the splitting effect caused by the image patch-based partition. We evaluate our framework using the benchmark dataset LPBA40. We demonstrate through our experimental result that our two patch-based strategies lead to better performance than the state-of-the-art techniques for unsupervised registration and segmentation.

2 Proposed Framework

In this section, we describe the overall workflow of our proposed framework.

Overview Workflow. In Fig. 1, our PC-SwinMorph first take the moving and fixed images as inputs. We then generate non-overlap patches from the two input images, and perform patch-level contrastive learning to refine the features (Patch-based Strategy I from Fig. 1). Then the contrasted features are fed into two weight-shared CNN encoders. Followed by a decoder, the features are recursively concatenated and enlarged with skip connections to reconstruct two sets of deformation field patches. We then use a 3D W-MSA and a 3D SW-MSA module [6] to refine and stitch the deformation field patches to obtain the full deformation field (Patch-based Strategy II from Fig. 1). Finally, we wrap the moving image → fixed image, and the fixed image → moving image. After the training registration process, we also adopt the full deformation field to transfer the segmentation mask for fixed masks to obtain the segmentation mask of the moving image. We underline that no masks are used in the training registration process, and they are only used in the testing segmentation stage. Hence, our framework is a unified unsupervised registration and segmentation network.

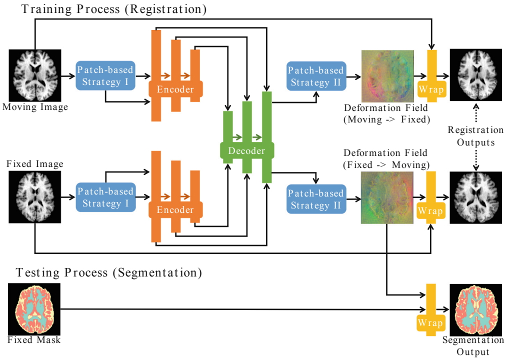


Fig. 1. Workflow of our proposed framework.

3 Experimental Results

In this section, we detail the experimental setup and experimental results to validate our proposed unified unsupervised registration and segmentation framework.

Experimental Setup. We evaluate our framework on the publicly available LONI Probabilistic Brain Atlas (LPBA40) dataset[1] using Dice evaluation metrics. For the implementation details regarding the network architecture, data pre-processing, and training and testing schemes, we refer to the reader to [5]. Our code will be publicly available upon the acceptance of this work.

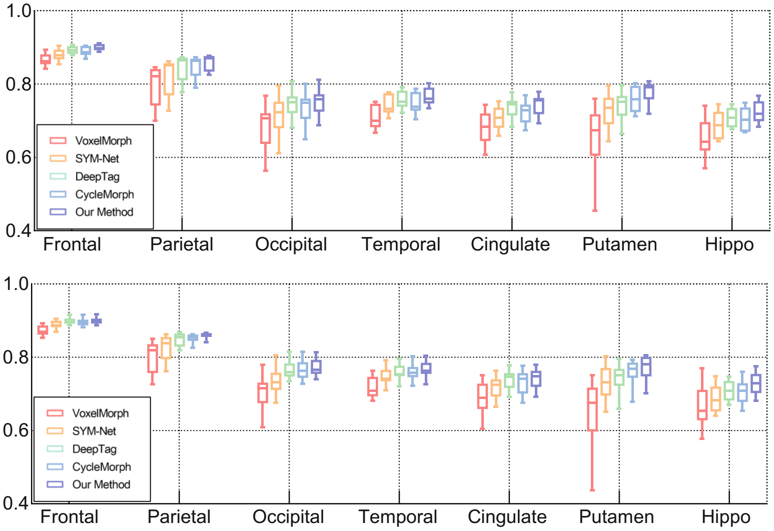


Fig. 2. Boxplots in terms of Dice, per anatomical region, for registration (top) and segmentation (bottom) tasks. The comparison displays our Method (PC-SwinMorph) against SOTA techniques.

Comparison to the State-of-the-Art Techniques. We compared our technique with recent unsupervised brain segmentation methods, including VoxelMorph [1], DeepTag [8], SYMNet [7], CycleMorph [2]. For a fair comparison, all models use the same backbone, VoxelMorph, which has been fine-tuned to achieve optimal performance. In Fig. 2, the boxplots summarise performance-wise, in terms of the *Dice* coefficient, the compared SOTA methods, and our PC-SwinMorph. In a closer look at the boxplots, we observe that our method

[1] https://loni.usc.edu/research/atlases.

outperforms all other SOTA methods by a large margin on all seven majority anatomical regions for both registration and segmentation tasks. Particularly, for both registration and segmentation tasks, our results report an improvement of 5.9% compared to VoxelMorph on the average Dice results, and 3.9–4.3% against the other compared SOTA techniques on the average Dice score.

4 Conclusion

We introduced a novel unified unsupervised framework for image registration and segmentation. We showed that patches are crucial for obtaining richer features and preserving anatomical details. Our intuition behind the performance gain of our technique, is that patches can capture not only global but also local spatial structures (more meaningful embeddings). We demonstrated, that at this point in time, our technique reported SOTA performance for both tasks.

References

1. Balakrishnan, G., et al.: Voxelmorph: a learning framework for deformable medical image registration. IEEE Trans. Med. Imaging **38**(8), 1788–1800 (2019)
2. Kim, B., et al.: Cyclemorph: cycle consistent unsupervised deformable image registration. Med. Image Anal. **71**, 102036 (2021)
3. Liu, L., Aviles-Rivero, A.I., Schönlieb, C.B.: Contrastive registration for unsupervised medical image segmentation. arXiv preprint arXiv:2011.08894 (2020)
4. Liu, L., Hu, X., Zhu, L., Heng, P.-A.: Probabilistic multilayer regularization network for unsupervised 3D brain image registration. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 346–354. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_39
5. Liu, L., Huang, Z., Liò, P., Schönlieb, C.B., Aviles-Rivero, A.I.: Pc-swinmorph: patch representation for unsupervised medical image registration and segmentation. arXiv preprint arXiv:2203.05684 (2022)
6. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. arXiv preprint arXiv:2103.14030 (2021)
7. Mok, T.C., Chung, A.: Fast symmetric diffeomorphic image registration with convolutional neural networks. In: CVPR, pp. 4644–4653 (2020)
8. Ye, M., et al.: Deeptag: an unsupervised deep learning method for motion tracking on cardiac tagging magnetic resonance images. In: CVPR, pp. 7261–7271 (2021)

Recent Developments of an Optimal Control Approach to Nonrigid Image Registration

Zicong Zhou[1](✉) and Guojun Liao[2]

[1] Institute of Natural Sciences, Shanghai Jiao Tong University, Shanghai, China
zicongzhou818@sjtu.edu.cn

[2] Math Department, University of Texas at Arlington, Arlington, TX, USA

Abstract. The Variational Principle (VP) forms diffeomorphisms (non-folding grids) with prescribed Jacobian determinant (JD) and curl under an optimal control set-up, which satisfies the properties of a Lie group. To take advantage of that, it is meaningful to regularize the resulting deformations of the image registration problem into the solution pool of VP. In this research note, (1) we provide an optimal control formulation of the image registration problem under a similar optimal control set-up as is VP; (2) numerical examples demonstrate the confirmation of diffeomorphic solutions as expected.

Keywords: Diffeomorphic image registration · Computational diffeomorphism · Jacobian determinant · Curl · *Green*'s identities

1 Our Approach to Image Registration

This work connects the resulting registration deformations to the solution pool of VP in [1], which achieves a recent progression in describing non-folding grids in a diffeomorphism group. Hence, to restrict the image registration method built in [3] satisfying the constraint of VP, it is reformulated and proposed as follows: let $\boldsymbol{I_m}$ be a ***moving*** image is to be registered to a ***fixed*** image $\boldsymbol{I_f}$ on the fixed and bounded domain $(\boldsymbol{\omega} =< x, y, z > \in)\Omega \subset \mathbb{R}^3$, the energy function *Loss* is minimized over the form $\boldsymbol{\phi} = \boldsymbol{id} + \boldsymbol{u}$ on Ω with $\boldsymbol{u} = \boldsymbol{0}$ on $\partial\Omega$,

$$Loss(\boldsymbol{\phi}) = \frac{1}{2}\int_{\Omega}[\boldsymbol{I_m}(\boldsymbol{\phi}) - \boldsymbol{I_f}]^2 d\boldsymbol{\omega} \quad \text{subjects to } \Delta\boldsymbol{\phi} = \boldsymbol{F}(f, \boldsymbol{g}) \text{ in } \Omega, \tag{1}$$

where the scalar-valued f and the vector-valued $\boldsymbol{g}$ are the control functions in the sense of VP that mimic the prescribed JD and curl, respectively.

A. Hering et al. (Eds.): WBIR 2022, LNCS 13386, pp. 194–197, 2022.
https://doi.org/10.1007/978-3-031-11203-4_22

1.1 Gradient with Respect to Control $\boldsymbol{F}$

The variational gradient of (1) with respect to $\delta\Delta\boldsymbol{\phi} = \delta\Delta\boldsymbol{u} = \delta\boldsymbol{F}$ is derived. For all $\delta\boldsymbol{F}$ vanishing on $\partial\Omega$ and by *Green*'s identities with fixed boundary condition,

$$\begin{aligned}\delta Loss(\boldsymbol{\phi}) &= \delta(\frac{1}{2}\int_\Omega [I_{\boldsymbol{m}}(\boldsymbol{\phi}) - I_{\boldsymbol{f}}]^2 d\boldsymbol{\omega}) = \int_\Omega [(I_{\boldsymbol{m}}(\boldsymbol{\phi}) - I_{\boldsymbol{f}})\nabla I_{\boldsymbol{m}}(\boldsymbol{\phi}) \cdot \delta\boldsymbol{\phi}]d\boldsymbol{\omega} \\ &= \int_\Omega [\Delta\boldsymbol{b} \cdot \delta\boldsymbol{\phi}]d\boldsymbol{\omega} = \int_\Omega [\boldsymbol{b} \cdot \delta\Delta\boldsymbol{\phi}]d\boldsymbol{\omega} = \int_\Omega [\boldsymbol{b} \cdot \delta\boldsymbol{F}]d\boldsymbol{\omega} \quad \Rightarrow \frac{\partial Loss}{\partial \boldsymbol{F}} = \boldsymbol{b},\end{aligned} \tag{2}$$

where $\Delta\boldsymbol{b} = (I_{\boldsymbol{m}}(\boldsymbol{\phi}) - I_{\boldsymbol{f}})\nabla I_{\boldsymbol{m}}(\boldsymbol{\phi})$, so, a gradient-based algorithm can be formed.

1.2 Hessian Matrix with Respect to Control Function $\boldsymbol{F}$

In case of a Newton optimizing scheme is applicable, from (2), one can derive the Hessian matrix $\boldsymbol{H}$ of (1) with respect to $\boldsymbol{F}$ as follows,

$$\delta^2 Loss(\boldsymbol{\phi}) := \delta(\delta Loss(\boldsymbol{\phi})) = \delta(\int_\Omega [(I_{\boldsymbol{m}}(\boldsymbol{\phi}) - I_{\boldsymbol{f}})\nabla I_{\boldsymbol{m}}(\boldsymbol{\phi}) \cdot \delta\boldsymbol{\phi}]d\boldsymbol{\omega}) = \int_\Omega [\delta\boldsymbol{\phi}^\top \boldsymbol{K}\delta\boldsymbol{\phi}]d\boldsymbol{\omega},$$

$$\text{where } \Delta^2\boldsymbol{H} = \boldsymbol{K} = \nabla I_{\boldsymbol{m}}(\boldsymbol{\phi})[\nabla I_{\boldsymbol{m}}(\boldsymbol{\phi})]^\top + (I_{\boldsymbol{m}}(\boldsymbol{\phi}) - I_{\boldsymbol{f}})\nabla^2 I_{\boldsymbol{m}}(\boldsymbol{\phi}),$$

$$\text{and } \nabla^2 I_{\boldsymbol{m}}(\boldsymbol{\phi}) = \begin{pmatrix} I_{\boldsymbol{m}}(\boldsymbol{\phi})_{xx} & I_{\boldsymbol{m}}(\boldsymbol{\phi})_{xy} & I_{\boldsymbol{m}}(\boldsymbol{\phi})_{xz} \\ I_{\boldsymbol{m}}(\boldsymbol{\phi})_{yx} & I_{\boldsymbol{m}}(\boldsymbol{\phi})_{yy} & I_{\boldsymbol{m}}(\boldsymbol{\phi})_{yz} \\ I_{\boldsymbol{m}}(\boldsymbol{\phi})_{zx} & I_{\boldsymbol{m}}(\boldsymbol{\phi})_{zy} & I_{\boldsymbol{m}}(\boldsymbol{\phi})_{zz} \end{pmatrix},$$

$$\text{so, } \delta^2 Loss(\boldsymbol{\phi}) = \int_\Omega [\delta\boldsymbol{\phi}^\top \Delta^2\boldsymbol{H}\delta\boldsymbol{\phi}]d\boldsymbol{\omega} = \int_\Omega [\delta\Delta\boldsymbol{\phi}^\top \boldsymbol{H}\delta\Delta\boldsymbol{\phi}]d\boldsymbol{\omega} \Rightarrow \frac{\partial^2 Loss}{(\partial\boldsymbol{F})^2} = \boldsymbol{H}. \tag{3}$$

A necessary condition that ensures a Newton scheme works is to show such Hessian $\boldsymbol{H}$ must be of Semi-Positive Definite matrix. This is left for future study.

1.3 Partial Gradients with Respect to Control Functions $\hat{\boldsymbol{f}}$ and $\boldsymbol{g}$

To ensure (1) producing diffeomorphic solutions that is controlled by $J_{min} \in (0, 1)$, instead of optimizing along $\boldsymbol{F}$ by (2), it can be set that $f := J_{min} + \hat{f}^2$ in (1). Since it is known $\delta\Delta\boldsymbol{u} = \delta\boldsymbol{F} = \delta(\nabla f - \nabla \times \boldsymbol{g})$, then, it carries to,

$$\begin{aligned}\delta Loss(\boldsymbol{\phi}) &= \int_\Omega [\boldsymbol{b} \cdot \delta\Delta\boldsymbol{\phi}]d\boldsymbol{\omega} = \int_\Omega [\boldsymbol{b} \cdot \delta\boldsymbol{F}]d\boldsymbol{\omega} = \int_\Omega [\boldsymbol{b} \cdot \delta(\nabla f - \nabla \times \boldsymbol{g})]d\boldsymbol{\omega} \\ &= \int_\Omega [\boldsymbol{b} \cdot (\nabla\delta(J_{min} + \hat{f}^2)]d\boldsymbol{\omega} + \int_\Omega [-\boldsymbol{b} \cdot \nabla \times \delta\boldsymbol{g}]d\boldsymbol{\omega}\end{aligned}$$

$$= \int_\Omega [\boldsymbol{b} \cdot (2\hat{f}\nabla\delta\hat{f})]d\boldsymbol{\omega} + \int_\Omega [-\boldsymbol{b} \cdot \nabla \times \delta\boldsymbol{g}]d\boldsymbol{\omega} = \int_\Omega [-2\hat{f}\nabla \cdot \boldsymbol{b}\delta\hat{f}]d\boldsymbol{\omega} + \int_\Omega [-\nabla \times \boldsymbol{b} \cdot \delta\boldsymbol{g}]d\boldsymbol{\omega}$$

$$\Rightarrow \quad \frac{\partial Loss}{\partial \hat{f}} = -2\hat{f}\nabla \cdot \boldsymbol{b} \quad \text{and} \quad \frac{\partial Loss}{\partial \boldsymbol{g}} = -\nabla \times \boldsymbol{b}. \tag{4}$$

2 Numerical Examples

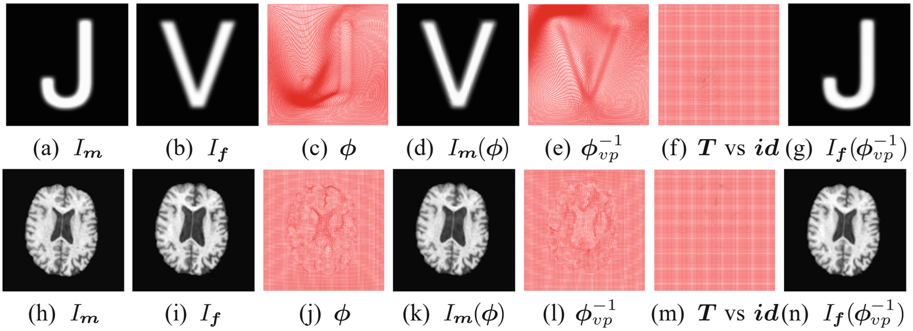
In our algorithms, $J_{min} = 0.5$ is artificially set. It is desirable to design a mechanism that yields optimal values of J_{min}. The gradient-based algorithms can be structured with (1) the coarse-to-fine **multiresolution** technique, which fits better in large deformation problems over binary images, as it did in [2]; and (2) the function composition **regriding** technique, which divides the problem difficulty and prevent non-diffeomorphic solutions on medical image registrations. These observations are demonstrated by the next example.

2.1 A Large Deformation Test and a MRI Registration Test

The J-to-V part of this example is done with **multiresolution** and the Brain Morph part is done with **regriding**. In Fig. 1(c, j), $\boldsymbol{\phi}$ is the diffeomorphic solution found by the proposed method; Fig. 1(d, k), $\boldsymbol{I_m}(\boldsymbol{\phi})$ is the registered image that is close to $\boldsymbol{I_f}$, Fig. 1(b, i). Next, $\boldsymbol{\phi}_{vp}^{-1}$ is the inverse of $\boldsymbol{\phi}$ that constructed by VP. In Fig. 1(f,m), $\boldsymbol{\phi}$ is composed by $\boldsymbol{\phi}^{-1}$, in Red grid, and superposed on Black grid $\boldsymbol{id}$ but the Black grid barely shows. This shows the composition $\boldsymbol{T} = \boldsymbol{\phi}_{vp}^{-1} \circ \boldsymbol{\phi}$ is very close to $\boldsymbol{id}$. Therefore, $\boldsymbol{\phi}_{vp}^{-1}$ can be treated as the inverse to $\boldsymbol{\phi}$ and they are of the same diffeomorphism group which VP focuses (Fig. 1).

Table 1. Evaluation of the proposed image registration

e.g.	Ω	$ratio = Loss(\boldsymbol{\phi})/Loss(\boldsymbol{id})$	$\min(\det\nabla\boldsymbol{\phi})$	JSC	DICE
`J-to-V`	$[1, 128]^2$	0.0034	0.2191	0.9337	0.9657
`Brain Morph`	$[1, 128]^2$	0.0605	0.2540	0.9849	0.9924


(a) I_m (b) I_f (c) ϕ (d) $I_m(\phi)$ (e) ϕ_{vp}^{-1} (f) $\boldsymbol{T}$ vs $\boldsymbol{id}$ (g) $I_f(\phi_{vp}^{-1})$

(h) I_m (i) I_f (j) ϕ (k) $I_m(\phi)$ (l) ϕ_{vp}^{-1} (m) $\boldsymbol{T}$ vs $\boldsymbol{id}$(n) $I_f(\phi_{vp}^{-1})$

Fig. 1. Resulting registration deformations and their inverses by VP

The question is whether $\boldsymbol{\phi}_{vp}^{-1}$ is also a valid inverse registration deformation that moves $\boldsymbol{I_f}$ back to $\boldsymbol{I_m}$. The answer is YES, at least in our tested examples. $\boldsymbol{I_f}(\boldsymbol{\phi}_{vp}^{-1})$

is indeed close to $\boldsymbol{I_m}$. That means $\boldsymbol{\phi}_{vp}^{-1}$ can be treated as a valid registration deformation from $\boldsymbol{I_f}$ to $\boldsymbol{I_m}$, as it is confirmed by the Table 2 records.

Table 2. Evaluation of $\boldsymbol{\phi}_{vp}^{-1}$ by VP in the sense of Image Registration

e.g.	*ratio* (of *Loss* from $\boldsymbol{I_f}(\boldsymbol{\phi}_{vp}^{-1})$ to $\boldsymbol{I_m}$)	$\min(\det\nabla\boldsymbol{\phi}_{vp}^{-1})$	JSC	DICE
`J-to-V`	0.0029	0.1520	0.9195	0.9581
`Brain morph`	0.0657	0.3212	0.9832	0.9915

3 Discussion

This note provides the analytic description with simple demonstration of the proposed method. A full paper with extensive experiments will be available soon.

References

1. Zhou, Z., Liao, G.: Construction of diffeomorphisms with prescribed jacobian determinant and curl. In: International Conference on Geometry and Graphics, Proceedings (2022). (in press)
2. Zhou, Z., Liao, G.: A novel approach to form Normal Distribution of Medical Image Segmentation based on multiple doctors′ annotations. In: Proceedings of SPIE 12032, Medical Imaging 2022: Image Processing, p. 1203237 (2022). https://doi.org/10.1117/12.2611973
3. Zhou, Z.: Image Analysis Based on Differential Operators with Applications to Brain MRIs, Ph.D. Dissertation, University of Texas at Arlington (2019)

2D/3D Quasi-Intramodal Registration of Quantitative Magnetic Resonance Images

Batool Abbas[1](✉), Riccardo Lattanzi[2], Catherine Petchprapa[2],
and Guido Gerig[1]

[1] Computer Science and Engineering, New York University Tandon School
of Engineering, New York, NY, USA
{batool.abbas,gerig}@nyu.edu

[2] Department of Radiology, New York University Grossman School of Medicine,
New York, NY, USA
riccardo.Lattanzi@nyulangone.org, catherine.petchprapa@nyumc.org

Abstract. Quantitative Magnetic Resonance Imaging (qMRI) is backed by extensive validation in research literature but has seen limited use in clinical practice because of long acquisition times, lack of standardization and no statistical models for analysis. Our research focuses on developing a novel quasi-intermodal 2D slice to 3D volumetric pipeline for an emerging qMR technology that aims to bridge the gap between research and practice. The two-part method first initializes the registration using a 3D reconstruction technique then refines it using a 3D to 2D projection technique. Intermediate results promise feasibility and efficacy of our proposed method.

1 Introduction

Biochemical changes often precede observable changes in morphologyand insight into these earlier asymptomatic deviations can help inform clinical strategy. Magnetic Resonance Imaging (MRI) has been traditionally used to acquire visual insight into the anatomy, morphology and physiology of living organisms. Quantitative MRI (qMRI) can capture and express the biochemical composition of the imaged structures as quantitative, calibrated physical units [8]. Despite a historically large body of research evidence providing validation for qMRI [3,17], it has seen limited integration into routine clinical practice due to obstacles such as infeasible acquisition time, insufficient standardization and a lack of statistical models for computational analysis [7].

One particularly promising approach for clinical integration of qMR enables rapid high-resolution and simultaneous mapping of multiple parameters in six 2D sections oriented around a central axis of rotation [10]. This method drastically cuts down acquisition time and has been proven to be highly reproducible [4] but it lacks normative models to perform comparative, population-based and longitudinal analysis. This is partly owing to the novelty of the data but also

A. Hering et al. (Eds.): WBIR 2022, LNCS 13386, pp. 198–205, 2022.
https://doi.org/10.1007/978-3-031-11203-4_23

because spatial normalization necessitates an effective 2D slice to 3D volume registration technique which continues to be an open problem today [5,12].

At first glance, this seems like a straightforward intermodal problem because the acquisition principle for both images is the same. However, the differences in protocol and parameters result in widely differing intensity distributions. The 3D qualitative volume comprises weighted intensity values while the 2D quantitative slices record raw un-weighted measurements. This difference categorizes this as a quasi-intra-modal registration problem [12] and adds another facet to its complexity. Thus the novelty of our proposed technology is rooted in both the originality of our data and in the research question that it aims to address.

1.1 Clinical Motivation

Symptomatic hip osteoarthritis (OA) is a degenerative joint disease that severely hinders functional mobility and impacts quality of life. It is one of the most common joint disorders in the United States [18] and the leading indication for primary total hip replacement surgeries [9]. The development of effective preventative and treatment measures necessitates the study of its causative factors.

1.2 Clinical Data

Each volunteer was scanned to collect a 3D qualitative scan of the hip and six 2D quantitative data scans acquired via incremental 30° rotations around a central axis passing through the femur bone as shown in Fig. 1. The specifics of the target 3D volume itself are less relevant since 3D/3D volume registration has several well-established and effective solutions that can be used to transform a template to a scan and vice versa [2]. For this reason, the 3D qualitative scan serves as our fixed volume. To start the process, a ‘localizer’ plane is maneuvered over the opening of the acetabulum by the MRI technician. The axis of rotation passes through this plane meaning that all acquired 2D scans are normal to this

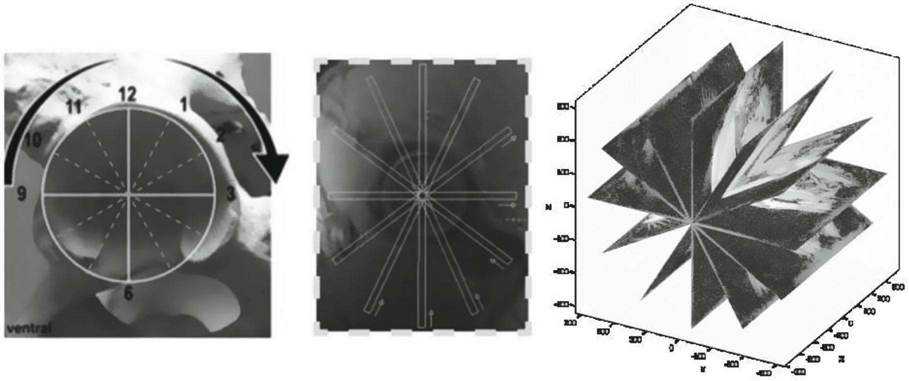


Fig. 1. Radial scans orientation(from left to right): i) superimposed over hip socket, ii) superimposed over a 2D MRI scan, iii) visualized in 3D space

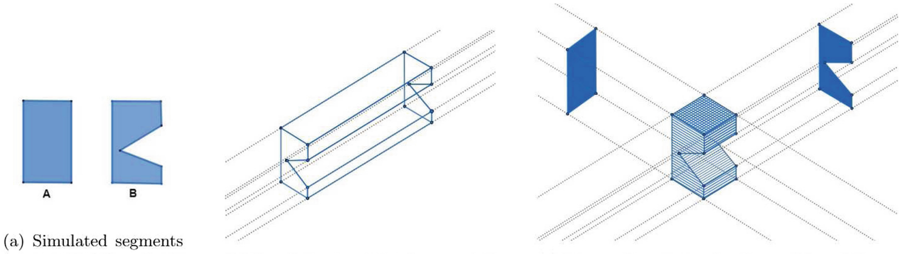
plane. The images in Fig. 1i) and ii) are parallel to this localizer. The mechanics of the acquisition technology [11] mean that only the center and normal vector of the localizer are accessible in the resulting DICOM images of all six qMR scans. The scans are expected roughly correlate to the diagram in Fig. 1i) but there is no guarantee with respect to the order or the directions of the final images. For comparison, a sequence of real scans from a volunteer can be compared to the expected orientations in the Appendix Fig. 4 and Fig. 5 respectively.

2 Method

Given the complexity of the anatomy imaged in these scans, we rely on initial segmentations of the femur and acetabulum to initialize our registrations. Different tissues express themselves differently in the modalities but the bony structures are consistently identifiable across all scans. We use a combination of random forests trained on samples from three different modalities of the 2D scans and a neural network pre-trained on a much larger dataset of shoulder joints to segment out the 2D and 3D bones respectively. Our proposed registration method can be broken down into two main steps. The first part includes recreating a visual hull [16] of the femur bone from the 2D slices that can then be registered to the 3D volume to estimate a reasonable initialization. The second step requires fine-tuning this registration through an iterative process of manipulation the 3D volume to 'emulate' the 2D slices, comparing these emulations to the real scans using a feature-based intermodal similarity metric such as mutual information [15] and updating the locations accordingly.

2.1 2D to 3D Reconstruction Using SFS

Shape-From-Silhouette (SFS) is a 3D reconstruction technique that uses images of 2D silhouettes to produce an output termed the visual or convex hull [16]. Traditional SFS problems are posed as a 3D object surrounded by cameras that capture 2D images of the object's silhouettes from their various point of views [6]. A simple model showing the moving parts of SFS can be seen in Appendix Fig. 3. In our case, the MR acquisition system can be reframed as an orthographic projection extending a polyhedral prism instead of a visual cone and with the bone segmentations as cross-sections. The six intersecting slices can be re-imagined as having been produced from similarly positioned external cameras surrounding the hip such that the 3D volume lies within the intersection of the visual prisms associated with the silhouette of the hip bones. This adaptation is illustrated in Fig. 2 and can be compared to the original SFS in Appendix Fig. 3.



(a) Simulated segments rotated 0° (A) and 90° (B)

(b) Parallel projection of segment B as silhouette defines a visual prism

(c) Intersection of visuals prisms of A and B create the visual hull containing the 3D object

Fig. 2. SFS reconstruction adapted to qMRI acquisition framework

Our exact reconstruction algorithm is still under construction, but we expect to share its details along with preliminary results in the next iteration of our publication. Additionally, this technique is based on the assumption that the relative locations of the 2D slices with respect to each other is accurate and known. This, as confirmed by Figs. 4 and 5, is not true for our case. To address this, we preprocessed our slices by comparing them to the expected orientations and manually aligned them into their appropriate positions. This process is also expected to be automated in the near future.

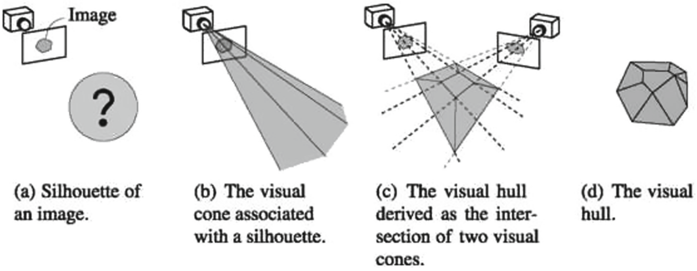
2.2 3D to 2D Projection Using Binary Search

Once the location of the 2D slices has been initialized, a set of 2D scans are 'emulated' via projection of the 3D volume onto the planes where the slices intersect. A second set of emulated scans are acquired after rotating these planes clockwise and anti-clockwise by an angle of 15°. The reason for this choice of angle is to explore the space of possibilities using a binary search in logarithmic time instead of an exhaustive linear search. All the actual 2D slices are then compared to these emulated slices and cumulatively vote to move the search space to one of the two sub-regions then repeat the process using 7° rotations and so on. Results from the binary search based optimization technique are pending the finalization of the initialization procedure, but results of the emulated scans from the 3D volumes can be seen in Appendix Figs. 6 and 7 respectively, visualized using 3DSlicer [1,13,14]

3 Discussion

Quantitative MRI technology allows earlier insight into asymptomatic morphological abnormalities that may improve the likelihood of positive prognoses. Despite extensive validation in research literature, qMRI has not translated into routine clinical practice for reasons including long acquisition times, lack of standardization and an absence of statistical models for analysis. Our research aims to enable a particularly promising new qMRI technology that has reduced acquisition times to a clinically feasible range and proven to be highly reproducible over time and scanners. Our contribution aims to enable registration of these 2D qMR slices to a normative 3D volumetric space to allow performing comparative and longitudinal analysis in larger scale or longer studies. We propose an initialization method using the 2D slices to create a 3D reconstruction and a followup optimization technique that emulates the qMRI acquisition process by capturing 2D slices from the 3D volume. While the method is currently under development, we have included intermediate results from its various sub-methods that show a lot of promise for the efficacy of our final 2D/3D multi-dimensional quasi-intermodal registration process.

Appendix



(a) Silhouette of an image.

(b) The visual cone associated with a silhouette.

(c) The visual hull derived as the intersection of two visual cones.

(d) The visual hull.

Fig. 3. Typical SFS reconstruction procedure as illustrated in [6] (a) Camera captures a silhouette (b) The silhouette defines a visual cone. (c) The intersection of two visual cones contains an object. (d) A visual hull of an object is the intersection of many visual cones

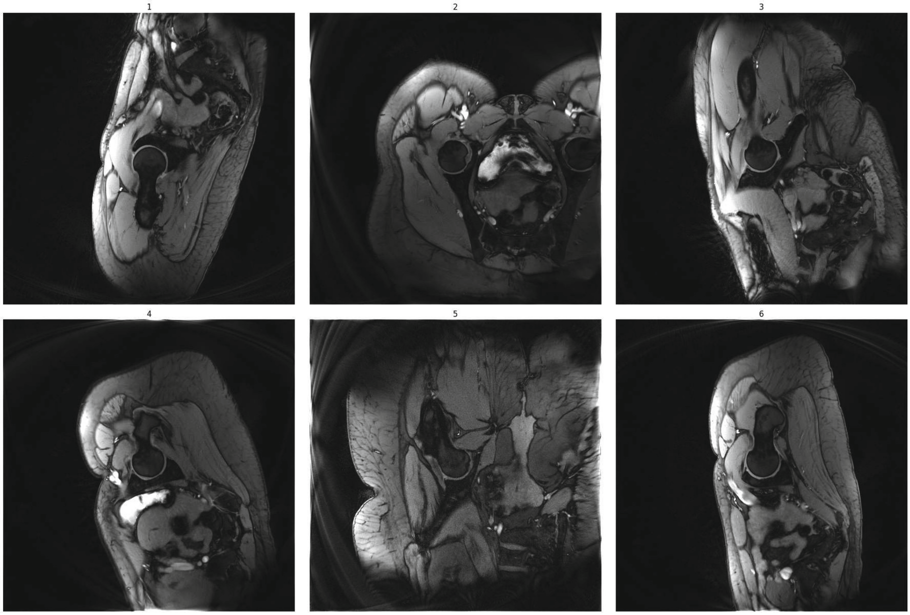


Fig. 4. Actual appearance of the six 2D qMR scans acquired from a volunteer

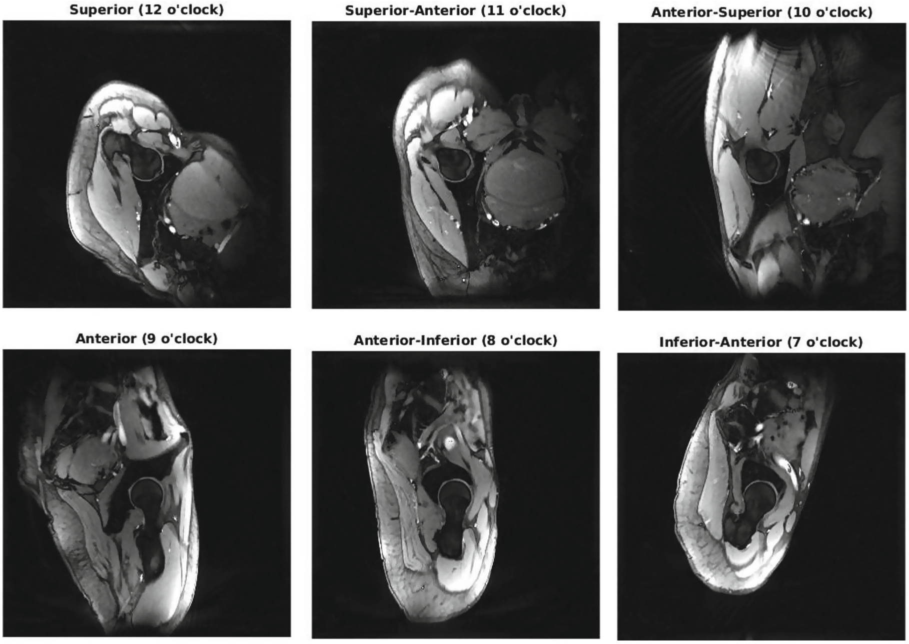

Superior (12 o'clock)
Superior-Anterior (11 o'clock)
Anterior-Superior (10 o'clock)
Anterior (9 o'clock)
Anterior-Inferior (8 o'clock)
Inferior-Anterior (7 o'clock)

Fig. 5. Expected appearance of the six 2D qMR scans

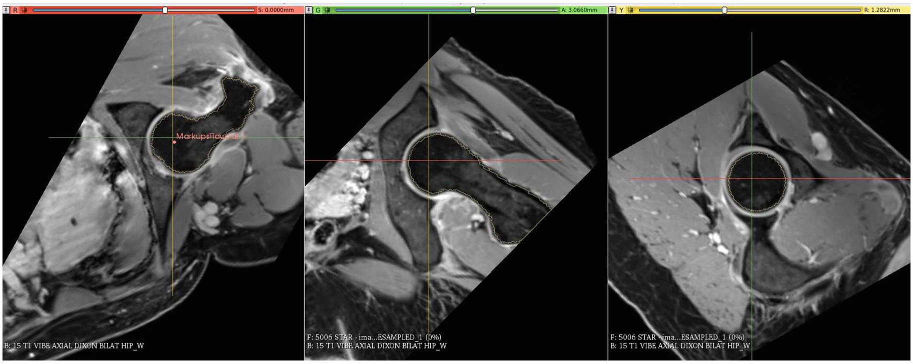

R
S: 0.0000mm
G
A: 3.0660mm
Y
R: 1.2822mm
MarkupsFiducial-1
B: 15 T1 VIBE AXIAL DIXON BILAT HIP_W
F: 5006 STAR - ima...ESAMPLED_1 (0%)
B: 15 T1 VIBE AXIAL DIXON BILAT HIP_W
F: 5006 STAR - ima...ESAMPLED_1 (0%)
B: 15 T1 VIBE AXIAL DIXON BILAT HIP_W

Fig. 6. The image on the far right depicts the localizer plane and so it is unchanging in both sets of emulations. The images on the center and left are captured using two planes orthogonal to the localizer plane and to each other. The initial locations of these planes was chosen arbitrarily but kept constant in both this and Fig. 7. This set produced via rotation by 90° clockwise

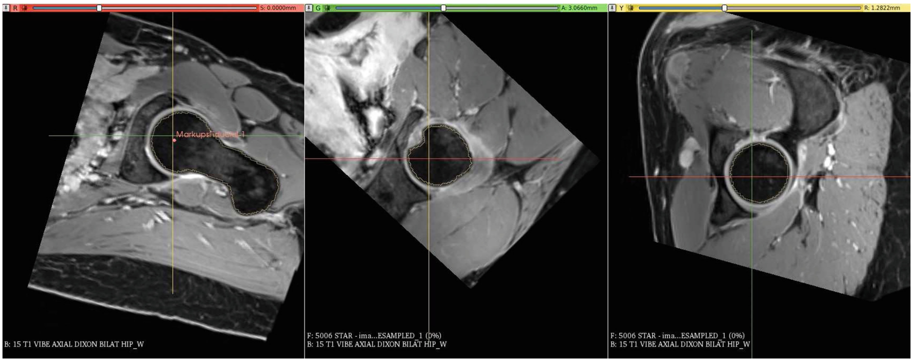

R
S: 0.0000mm
G
A: 3.0660mm
Y
R: 1.2822mm
MarkupsFiducial-1
B: 15 T1 VIBE AXIAL DIXON BILAT HIP_W
F: 5006 STAR - ima...ESAMPLED_1 (0%)
B: 15 T1 VIBE AXIAL DIXON BILAT HIP_W
F: 5006 STAR - ima...ESAMPLED_1 (0%)
B: 15 T1 VIBE AXIAL DIXON BILAT HIP_W

Fig. 7. The image on the far right depicts the localizer plane and so it is unchanging in both sets of emulations. The images on the center and left are captured using two planes orthogonal to the localizer plane and to each other. The initial locations of these planes was chosen arbitrarily but kept constant in both this and Fig. 6. This set produced via rotation by 45° counter-clockwise

References

1. Fedorov, A., et al.: 3D Slicer as an image computing platform for the Quantitative Imaging Network. Magn. Reson. Imaging **30**(9), 1323–1341 (2012). https://www.slicer.org/
2. Avants, B.B., Tustison, N.J., Stauffer, M., Song, G., Wu, B., Gee, J.C.: The insight toolkit image registration framework. Front. Neuroinf. **8**, 44 (2014)
3. Bashir, A., Gray, M.L., Burstein, D.: Gd-dtpa2- as a measure of cartilage degradation. Magn. Reson. Med. **36**, 665–673 (1996)

4. Cloos, M.A., Assländer, J., Abbas, B., Fishbaugh, J., Babb, J.S., Gerig, G., Lattanzi, R.: Rapid radial t1 and t2 mapping of the hip articular cartilage with magnetic resonance fingerprinting. J. Magn. Reson. Imaging **50**(3), 810–815 (2019)
5. Ferrante, E., Paragios, N.: Slice-to-volume medical image registration: a survey. Med. Image Anal. **39**, 101–123 (2017)
6. Imiya, A., Sato, K.: Shape from silhouettes in discrete space. In: Gagalowicz, A., Philips, W. (eds.) CAIP 2005. LNCS, vol. 3691, pp. 296–303. Springer, Heidelberg (2005). https://doi.org/10.1007/11556121_37
7. Jazrawi, L.M., Alaia, M.J., Chang, G., Fitzgerald, E.F., Recht, M.P.: Advances in magnetic resonance imaging of articular cartilage. J. Am. Acad. Orthopaedic Surg. **19**, 420–429 (2011)
8. Jazrawi, L.M., Bansal, A.: Biochemical-based MRI in diagnosis of early osteoarthritis. Imaging Med. **4**(1), 01 (2012)
9. Katz, J.N., et al.: Association between hospital and surgeon procedure volume and outcomes of total hip replacement in the united states medicare population. JBJS **83**(11), 1622–1629 (2001)
10. Lattanzi, R., et al.: Detection of cartilage damage in femoroacetabular impingement with standardized dgemric at 3 t. Osteoarthritis Cartilage **22**(3), 447–456 (2014)
11. Ma, D., et al.: Magnetic resonance fingerprinting. Nature **495**(7440), 187 (2013)
12. Markelj, P., Tomaževič, D., Likar, B., Pernuš, F.: A review of 3D/2D registration methods for image-guided interventions. Med. Image Anal. **16**(3), 642–661 (2012). https://doi.org/10.1016/j.media.2010.03.005, https://www.sciencedirect.com/science/article/pii/S1361841510000368, computer Assisted Interventions
13. Pieper, S., Halle, M., Kikinis, R.: 3D slicer. In: 2004 2nd IEEE International Symposium on Biomedical Imaging: Nano to Macro (IEEE Cat No. 04EX821), pp. 632–635. IEEE (2004)
14. Pieper, S., Lorensen, B., Schroeder, W., Kikinis, R.: The NA-MIC kit: ITK, VTK, pipelines, grids and 3D slicer as an open platform for the medical image computing community. In: 3rd IEEE International Symposium on Biomedical Imaging: Nano to Macro, 2006, pp. 698–701. IEEE (2006)
15. Pluim, J.P., Maintz, J.A., Viergever, M.A.: Mutual-information-based registration of medical images: a survey. IEEE Trans. Med. Imaging **22**(8), 986–1004 (2003)
16. Schneider, D.C.: Shape from silhouette, pp. 725–726. Springer, Boston (2014). https://doi.org/10.1007/978-0-387-31439-6_206
17. Venn, M., Maroudas, A.: Chemical composition and swelling of normal and osteoarthrotic femoral head cartilage. I. Chemical composition. Ann. Rheumatic Dis. **36**, 121–129 (1977)
18. Zhang, Y., Jordan, J.M.: Epidemiology of osteoarthritis. Clin. Geriatric Med. **26**(3), 355–369 (2010)

Deep Learning-Based Longitudinal Intra-subject Registration of Pediatric Brain MR Images

Andjela Dimitrijevic[1,2](✉), Vincent Noblet[3](✉), and Benjamin De Leener[1,2,4](✉)

[1] NeuroPoly Lab, Institute of Biomedical Engineering, Polytechnique Montréal, Montréal, QC, Canada
{andjela.dimitrijevic,benjamin.de-leener}@polymtl.ca

[2] Research Center, Ste-Justine Hospital University Centre, Montréal, QC, Canada

[3] ICube-UMR 7357, Université de Strasbourg, CNRS, Strasbourg, France
vincent.noblet@unistra.fr

[4] Computer Engineering and Software Engineering, Polytechnique Montréal, Montréal, QC, Canada

Abstract. Deep learning (DL) techniques have the potential of allowing fast deformable registration tasks. Studies around registration often focus on adult populations, while there is a need for pediatric research where less data and studies are being produced. In this work, we investigate the potential of unsupervised DL-based registration in the context of longitudinal intra-subject registration on 434 pairs of publicly available Calgary Preschool dataset of children aged 2–7 years. This deformable registration task was implemented using the DeepReg toolkit. It was tested in terms of input spatial image resolution (1.5 vs 2.0 mm isotropic) and three pre-alignement strategies: without (NR), with rigid (RR) and with rigid-affine (RAR) initializations. The evaluation compares regions of overlap between warped and original tissue segmentations using the Dice score. As expected, RAR with an input spatial resolution of 1.5 mm shows the best performances. Indeed, RAR has an average Dice score of of 0.937 ± 0.034 for white matter (WM) and 0.959 ± 0.020 for gray matter (GM) as well as showing small median percentages of negative Jacobian determinant (JD) values. Hence, this shows promising performances in the pediatric context including potential neurodevelopmental studies.

Keywords: Learning-based image registration · Pediatric · MRI

1 Introduction

Registration consists of bringing a pair of images into spatial correspondence. There are hardly any registration methods dedicated to the pediatric brain,

Supported by Polytechnique Montréal, by the Canada First Research Excellence Fund, and by the TransMedTech Institute.

A. Hering et al. (Eds.): WBIR 2022, LNCS 13386, pp. 206–210, 2022.
https://doi.org/10.1007/978-3-031-11203-4_24

mainly because of the difficulties arising from major changes that occur during neurodevelopment [5]. Conventional deformable registration involves estimating a deformation field through an iterative optimization problem. This process is time consuming, but provides accurate results. Convolutional neural networks (CNN) can allow faster registrations by applying a learning-based approach [4]. Hence, applying DL methods to pediatric brain scans could improve registration and future diagnostics for medical applications. Ultimately, it would be relevant to validate the potential use of DL-based frameworks for pediatric populations.

The general objective of this study is to validate a DL framework which allows fast intra-subject deformable registrations after training on pediatric MRI scans. To do so, different initial conditions are considered by fragmenting the non-rigid transformation into its simpler parts. Pre-network rigid registration, RigidReg (RR) and rigid-affine registration, RigidAffineReg (RAR) are performed on each intra-subject pair using ANTs [1] in order to determine their respective impact on the network's performance. Also, a third method called NoReg (NR) is investigated where no pre-alignment task is done. These three methods are then trained using a U-Net like CNN architecture implemented via the DeepReg toolkit [3]. The robustness of these DL techniques is assessed by using different input resolutions (1.5 vs 2.0 mm isotropic) for the same network architectures.

2 Methodology

Preprocessing Pipeline. Each image was corrected for bias field inhomogeneity using N4 algorithm. Both rigid and rigid-affine pre-alignments were performed with ANTs registration framework. The Mattes similarity metric was used.

Unsupervised Deformable Registration Framework. The U-Net architecture used to generate the deformation field consists of a 3-layer encoder and decoder with 8, 16 and 32 channels each. As for the loss function, it is composed of a local normalized cross-correlation similarity measure and an L2-norm gradient regularization factor to ensure realistic physical deformation fields. Local normalized cross-correlation is chosen for its robustness to local variations of intensities. The ADAM optimizer is used with a learning rate set to 1.0e-4. Finally, the network was trained on a GeForce RTX 2080 Ti GPU.

3 Experiments

Data. 434 pairs of moving/fixed 3D images were extracted from the longitudinal Calgary Preschool dataset [6] containing 247 T1-weighted images from 64 children aged 2–7 years old. The average time interval between consecutive scans is of 1.15 ± 0.68 years. The original images have a native resolution of $0.4492 \times 0.4492 \times 0.9$ mm^3. The resized images of 1.5 mm as well as 2.0 mm isotropic resolution have respectively a matrix size of $153 \times 153 \times 125$ and $114 \times 114 \times 94$.

Evaluation. To acquire white matter (WM), gray matter (GM) and cerebrospinal fluid (CSF) segmentations for evaluation purposes, each image was non-linearly registered to the MNI pediatric template for children 4.5–8.5 years old [2]. This also allowed obtaining skull-stripped images via the available mask in the template space. Unsupervised networks are then evaluated using the Dice score as a performance metric. In addition, the generated deformation fields are evaluated using the percentage of negative JD values indicating unwanted local foldings. To compare the impact from the three initialization methods or input resolutions, one-sided Wilcoxon signed-rank tests were performed.

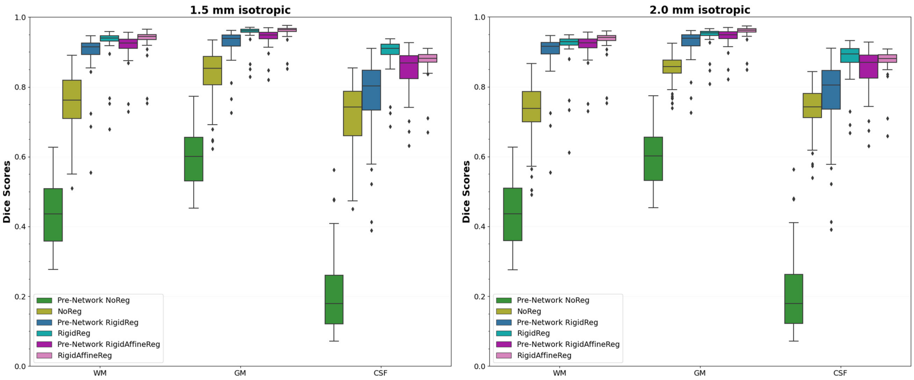


Fig. 1. Dice scores results for different input resolutions obtained for each method compared to their pre-network Dice scores represented as boxplots. The Dice scores are calculated for all subjects and WM, GM and CSF regions using the test set.

Table 1. Average Dice scores per resolution calculated over all segmented regions and subjects using the test set for the three studied methods. Median percentages of negative JD values are given because of highly right-skewed distributed data. ANTs pre-registration tasks are performed on the native resolution of 0.4492 × 0.4492 × 0.9 mm^3 and using the available CPU implementation.

Methods	1.5 mm isotropic				2.0 mm isotropic				Native resolution
	Dice score	% of JD<0	Train time/epoch	Test time/pair	Dice score	% of JD<0	Train time/epoch	Test time/pair	ANTs pre-reg time/pair
NR	0.764 ± 0.105	1.11e–1	189.3 s	4.88 s	0.770 ± 0.088	1.33e–1	74.8 s	1.87 s	0 s
RR	0.929 ± 0.045	1.86e–4	137.7 s	3.55 s	0.916 ± 0.051	0	78.1 s	1.88 s	168.6 s
RAR	0.924 ± 0.047	0	177.5 s	4.13 s	0.922 ± 0.047	0	75.8 s	1.86 s	365.8 s

4 Results

A 85/15 % split was respectively done for train and test sets. Then, a three-fold cross-validation technique is employed to train and evaluate each method containing 123 pairs per fold. In total, 65 pairs are used for test purposes. Above, presented results for the two evaluated resolutions in Table 1 and Fig. 1 come from this unseen test set. Figure 2 shows the differences of obtained predicted

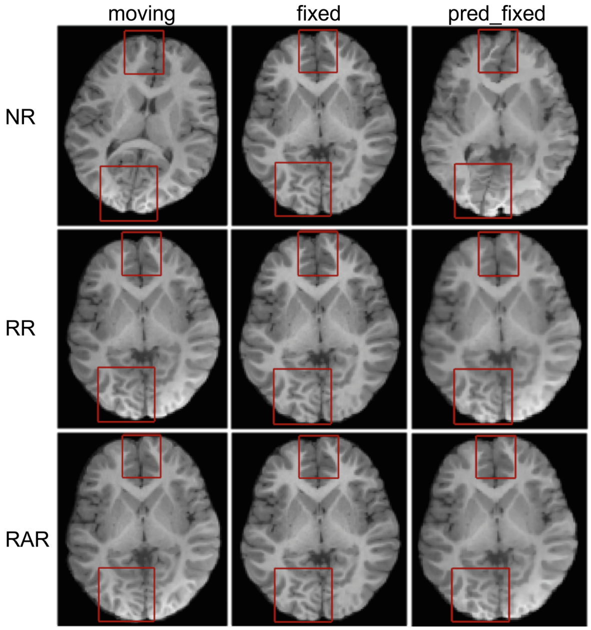


Fig. 2. Resulting images for a specific pair (age interval of 3.37 years) for the three pre-alignment strategies using an input spatial image resolution of 1.5 mm isotropic.

fixed images for all the three considered initialization methods. Also, two one-sided Wilcoxon tests were conducted comparing, first, the median Dice scores differences between RR and RAR (RAR-RR) for both resolutions as well as between 2.0 and 1.5 (1.5–2.0) for all initialization methods. The first test allowed rejecting the null hypothesis only for WM and GM ($p<1.06e{-}10$) showing higher median Dice scores for RAR compared to RR for both resolutions. The second test shows that 1.5 mm isotropic resolution, at the cost of longer train and test times, yields slightly, but statistically significant better performances than 2.0 for RR and RAR methods for all segmented regions ($p<1.51e{-}4$). This improvement is not significant for GM and CSF regions for the NR method.

5 Discussion and Conclusion

In this study, we demonstrated that DL-based deformable registration succeeds to improve registration accuracy regardless of the initialization method and for both tested resolutions (see Fig. 1). RAR demonstrated higher Dice scores compared to RR for WM (0.937 ± 0.034 vs 0.930 ± 0.046) and GM (0.959 ± 0.020 vs 0.955 ± 0.025). Differing results for CSF may be due to its thin surface and errors arising from the skull-stripping process. Both RR and RAR reached high registration quality, while NR shows lower registration performance due to its incapacity to extract both global and local transformations simultaneously, shown in

Fig. 2. However, NR remains relevant as no prior registration is needed. Future work will evaluate the capacity of a neural network to decompose the global and local transformations. Considering all studied combinations of pre-alignment strategies and input resolutions, RAR provides better Dice scores with 1.5 mm isotropic resolution images, which could help to perceive neurodevelopmental changes from a large age range of pediatric data.

References

1. Avants, B.B., Tustison, N.J., Song, G., Cook, P.A., Klein, A., Gee, J.C.: A reproducible evaluation of ants similarity metric performance in brain image registration. NeuroImage **54**(3), 2033–2044 (2011). https://doi.org/10.1016/j.neuroimage.2010.09.025, https://www.sciencedirect.com/science/article/pii/S1053811910012061
2. Fonov, V., Evans, A.C., Botteron, K., Almli, C.R., McKinstry, R.C., Collins, D.L.: Unbiased average age-appropriate atlases for pediatric studies. NeuroImage **54**(1), 313–327 (2011). https://doi.org/10.1016/j.neuroimage.2010.07.033, https://www.sciencedirect.com/science/article/pii/S1053811910010062
3. Fu, Y., et al.: Deepreg: a deep learning toolkit for medical image registration. J. Open Source Softw. **5**(55) (2020). https://doi.org/10.21105/joss.02705
4. Haskins, G., Kruger, U., Yan, P.: Deep learning in medical image registration: a survey. Mach. Vision Appl. **31**(1–2) (2020). https://doi.org/10.1007/s00138-020-01060-x
5. Phan, T.V., Smeets, D., Talcott, J.B., Vandermosten, M.: Processing of structural neuroimaging data in young children: bridging the gap between current practice and state-of-the-art methods. Dev. Cogn. Neurosci. **33**, 206–223 (2018). https://doi.org/10.1016/j.dcn.2017.08.009
6. Reynolds, J.E., Long, X., Paniukov, D., Bagshawe, M., Lebel, C.: Calgary preschool magnetic resonance imaging (MRI) dataset. Data Brief **29**, 105224 (2020). https://doi.org/10.1016/j.dib.2020.105224, https://www.ncbi.nlm.nih.gov/pubmed/32071993

Real-Time Alignment for Connectomics

Neha Goyal(✉), Yahiya Hussain, Gianna G. Yang, and Daniel Haehn

University of Massachusetts - Boston, Boston, MA 02125, USA
sneh.goyal.22@gmail.com

Abstract. In Connectomics, researchers are creating the brain's wiring diagram at nanometer resolution. As part of this processing workflow, 2D electron microscopy (EM) images must be aligned to 3D volumes. However, existing alignment methods are computationally expensive and can take a long time. We hypothesize that adding biological features improve and accelerate the alignment procedure. Since especially mitochondria can be detected accurately and fast, we propose a new alignment method, MITO, that uses these structures as landmark points. With MITO, we can decrease the alignment time by 27%, and our experiments indicate a throughput of 33 Megapixels/s, which is faster than the acquisition speed of current microscopes. We can align an image volume of 1268 × 1524 × 160 voxels in less than 12 s. We compare our method to the following feature generators: ORB, BRISK, FAST, and FREAK.

Keywords: Image alignment · Registration · Feature matching

1 Introduction

Connectomics studies the functional and structural connections of a brain to understand the correlation between the physiology of the brain and its behavior. This correlation will help better treatment solutions, design new drugs for mental pathologies, construct custom neural prostheses, etc. Therefore, a registration process is required to map every synaptic connection to build a computer-generated brain wiring diagram. When needed, the image registration process is necessary to map the similarities between images acquired at different times or across other subjects by various sensors. Moreover, image registration is a crucial processing step in various other bio-medical image applications. In this study, we used diamond-knife-sliced electron microscopy (EM) images that provide high resolution such that individual synaptic connections between neurons are visible. We hypothesize to align these images by adding biological features can improve state-of-the-art registration methods. We have used a feature extraction model that follows four steps: feature detection, feature extraction, feature matching, and estimating the transformation matrix. Using the biological features, we get faster real-time alignment performance.

A. Hering et al. (Eds.): WBIR 2022, LNCS 13386, pp. 211–214, 2022.
https://doi.org/10.1007/978-3-031-11203-4_25

2 Methods

We used unaligned two-dimensional EM images with nanometer resolution, and the corresponding mitochondria mask data as labeled data. The original dataset is called Lucchi++ and was the result of the study 'Fast Mitochondria Detection for Connectomics [1].' This dataset included two stacks: image and mask of 160 tiles, each having 768 × 1024 px. We created the unaligned dataset from the original by rotating each image tile and its corresponding mask tile at an arbitrary angle between $(-\pi, +\pi)$ and added a pad size of 250 px on all the sides to prevent information loss at the time of rotation. The new unaligned dataset has two stacks: image and mask, with 160 tiles and dimensions 1268 × 1524 px.

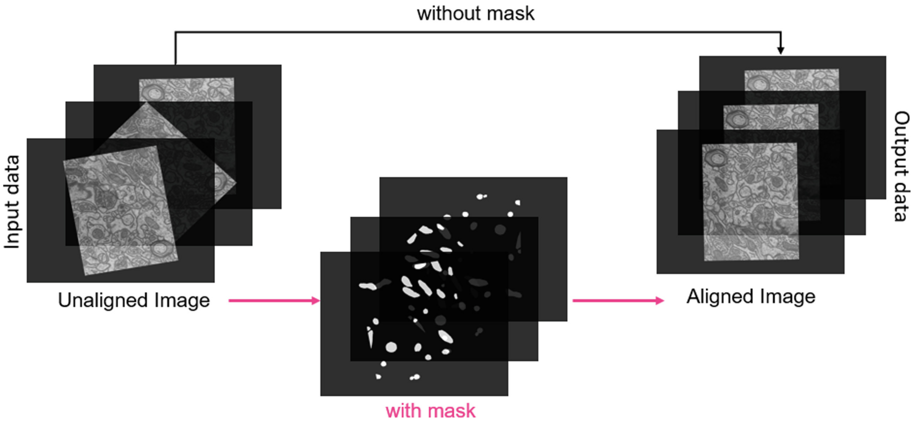


Fig. 1. Mapping of input images with and without adding the biological features. The unaligned input EM images (left) were mapped in real-time with and without adding the biological features (mask data). We generated a stack of aligned images (right) as output in both the cases to draw comparisons.

We performed an automatic registration on the unaligned EM images using a custom-build interactive program that runs the feature extraction model and calculates alignment score, execution time, and throughput for the entire dataset. This model used existing computer vision algorithms such as FAST [6], ORB [2], BRISK [3] to learn the features or patterns from the input dataset. We propose a new feature detector mechanism called **MITO** that detects the keypoints in EM images using mitochondria from mask images as a region of interest (ROI). In this feature detection step, we introduced mask images as additional biological features to improve the alignment performance. In the feature description step, the model uses ORB, BRISK, and FREAK [4] algorithms to create descriptors that are unique and could be referred to as a keypoint's numerical fingerprint. In the next step, we used feature matching algorithms such as BF [8] and FLANN [9] matcher to map (x_i, y_i) of the source image to $(x_i^{'}, y_i^{'})$ of the target image. Finally, with the help of the homography matrix, the model transforms the source image and outputs the aligned image. We generated two stacks of registered images with and without the help of mitochondria masks for comparisons (see Fig. 1).

3 Results

We perform experiments on the unaligned Lucchi++ dataset to measure timing and alignment accuracy. When we combine biological features using the MITO method with the BF and FLANN matchers, we observe a maximum execution time of 9.49 (±0.37) seconds for the whole stack. When comparing the accuracy, we measure a dice score of over 0.89 for both BF and FLANN, indicating quality alignment. The average throughput with MITO is at least 33 Megapixels/s which is faster than the acquisition speed of modern electron microscopes (11 Megapixels/s). Our findings indicate that MITO can be used to align connectomics image data in real-time during image acquisition. Table 1 shows the full evaluation.

Table 1. Alignment Results on Lucchi++. We compare the BF and FLANN matchers with a variety of feature descriptors. When using the MITO detector, we measure the throughput of at least 33 Megapixels/s, indicating real-time performance.

Matcher	Detector + Descriptor	Mask	Dice score	Execution time (sec.)	Stack throughput (MP/s)
BF	BRISK	✓	0.9354	47.0052(±1.5173)	6.7879(±0.2170)
			0.8569	**19.3020(±0.2625)**	**16.5210(±0.2256)**
	ORB	✓	0.7529	19.4427(±1.8462)	16.4941(±1.4953)
			0.8226	20.4218(±0.5493)	15.6208(±0.4259)
	FAST + BRISK	✓	0.9184	2419.9270(±99.9857)	0.1319(±0.0053)
			0.8762	**28.4635(±1.2776)**	**11.2167(±0.4908)**
	ORB + BRISK	✓	0.6291	16.3020(±1.4923)	19.6693(±1.8124)
			0.7935	16.9687(±1.6858)	18.9180(±1.9290)
	FAST + FREAK	✓	0.9405	2391.9479(±137.7484)	0.1335(±0.0074)
			0.9140	**25.1302(±0.5)**	**12.6912(±0.2498)**
	ORB + FREAK	✓	0.8320	16.6458(±1.8088)	19.2979(±1.9733)
			0.7637	16.8072(±0.1365)	18.9718(±0.1545)
	MITO(ours) + BRISK	✓	**0.9142**	**7.7708(±0.0888)**	**41.035(±0.4713)**
	MITO(ours) + FREAK	✓	**0.8963**	**8.3697(±0.0888)**	**38.0983(±0.4027)**
FLANN	BRISK	✓	0.9344	40.1145(±0.9393)	7.9514(±0.1887)
			0.8338	**19(±2.4111)**	**16.9513(±2.0058)**
	ORB	✓	0.8069	19.3802(±1.2145)	16.4941(±0.9979)
			0.8280	20.6875(±1.1149)	15.4417(±0.8082)
	FAST + BRISK	✓	0.9338	3082.2343(±130.2627)	0.1035(±0.0043)
			0.8784	**29.6041(±0.2350)**	10.7709(±0.0856)
	ORB + BRISK	✓	0.6297	16.9322(±1.7772)	18.9655(±1.9261)
			0.7648	15.2031(±1.1735)	21.0579(±1.6571)
	FAST + FREAK	✓	0.9450	2628.3229(±32.5343)	0.1213(±0.0015)
			0.9091	**31.4166(±4.7502)**	**10.2940(±1.4380)**
	ORB + FREAK	✓	0.8285	16.2812(±0.0563)	19.5841(±0.0676)
			0.7402	17.2083(±1.2107)	18.5882(±1.2665)
	MITO(ours) + BRISK	✓	**0.9062**	**9.2239(±0.7265)**	**34.7050(±2.6154)**
	MITO(ours) + FREAK	✓	**0.8928**	**9.4843(±0.3694)**	**33.6528(±1.3213)**

4 Conclusion

Fast registration is crucial to creating 3D volumetric connectomics datasets from unaligned EM images. This process can be computationally expensive. Based on our studies, adding biological features to register these images results in faster alignment. Specifically, we include mitochondria masks as part of our MITO feature detector. With MITO, the overall dice score is higher than 0.80, and the throughput is faster than 11 Megapixels/s. These measurements indicate the possibility of real-time alignment during the image acquisition with modern electron microscopes.

References

1. Casser, V., Kang, K., Pfister, H., Haehn, D.: Fast mitochondria detection for connectomics. In: Proceedings of the Third Conference on Medical Imaging with Deep Learning, PMLR, pp. 111–120 (2020)
2. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.R.: ORB: an efficient alternative to SIFT or SURF. In: 2011 International Conference on Computer Vision, pp. 2564–2571. IEEE (2011)
3. Leutenegge, S., Chli, M., Siegwart, R.Y.: BRISK: binary robust invariant scalable keypoints. In: 2011 International Conference on Computer Vision, pp. 2548–2555. IEEE (2011)
4. Alahi, A., Ortiz, R., Vandergheynst, P.: Freak: fast retina keypoint. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 510–517 (2012)
5. Calonder, M., Lepetit, V., Strecha, C., Fua, P.: BRIEF: binary robust independent elementary features. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6314, pp. 778–792. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15561-1_56
6. Rosten, E., Drummond, T.: Machine learning for high-speed corner detection. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3951, pp. 430–443. Springer, Heidelberg (2006). https://doi.org/10.1007/11744023_34
7. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM **24**, 381–395 (1981)
8. OpenCV modules. http://docs.opencv.org/3.1.0. Accessed 19 Apr 2017
9. Muja, M., Lowe, D.G.: Fast approximate nearest neighbors with automatic algorithm configuration. In: VISAPP (2009)
10. Khachikian, S., Emadi, M.: Applying fast & freak algorithms in selected object tracking. Int. J. Adv. Res. Electr. Electron. Instrument. Eng. **5**, 5829–5839 (2016)
11. Phan, D., Oh, C.-M., Kim, S.-H., Na, I.-S., Lee, C.-W.: Object recognition by combining binary local invariant features and color histogram. In: 2013 2nd IAPR Asian Conference on Pattern Recognition (ACPR), pp. 466–470 (2013)
12. Rosten, E., Porter, R., Drummond, T.: Faster and better: a machine learning approach to corner detection. IEEE Trans. Pattern Anal. Mach. Intell. **32**, 105–119 (2010)
13. Szeliski, R.: Computer Vision: Algorithms and Applications. Springer-Verlag, London (2010). https://doi.org/10.1007/978-1-84882-935-0
14. Li, X., Zhanyi, H.: Rejecting mismatches by correspondence function. Int. J. Comput. Vision **89**, 1–17 (2010)

Weak Bounding Box Supervision for Image Registration Networks

Mona Schumacher[1,2](✉), Hanna Siebert[1], Ragnar Bade[2], Andreas Genz[2], and Mattias Heinrich[1]

[1] Institute of Medical Informatics, University of Luebeck, Luebeck, Germany
[2] MeVis Medical Solutions AG, Bremen, Germany
mona.schumacher@mevis.de

Abstract. Image registration is a fundamental task in medical image analysis. Many deep learning based methods use multi-label image segmentations during training to reach the performance of conventional algorithms. But the creation of detailed annotations is very time-consuming and expert knowledge is essential. To avoid this, we propose a weakly supervised learning scheme for deformable image registration that uses bounding boxes during training. By calculating the loss function based on these bounding box labels, we are able to perform an image registration with large deformations without using densely labeled annotations. The performance of the registration of inter-patient 3D Abdominal CT images can be enhanced by approximately 10% only with little annotation effort in comparison to unsupervised learning methods. Taken into account this annotation effort, the performance also exceeds the performance of the label supervised training.

Keywords: Deformable image registration · Weak supervision · Bounding box supervision

1 Introduction

Medical image registration is the process of the alignment of the anatomical structures of two or more images in order to be able to do follow up studies, image-guidance or to plan a treatment. Deep learning methods have become increasingly important. They have demonstrated low computation times and are promising to enable real time registration approaches. For the case of brain image registration [1], which only require small deformation, already satisfactory results could be achieved. The registration of images of highly deformable body regions, such as the abdominal region or thorax are, due to the respiration or digestion, more complex and still often solved with conventional algorithms [2,3]. Deep learning methods have started to address the challenge of handling large deformations (for example in the Learn2Reg Challenge, cf. learn2reg.grand-challenge.org) [4,5]. Mok et al. [6] use Laplacian pyramids to solve the registration in a coarse-to-fine scheme inspired by classical algorithms. They show that

A. Hering et al. (Eds.): WBIR 2022, LNCS 13386, pp. 215–219, 2022.
https://doi.org/10.1007/978-3-031-11203-4_26

label supervision substantially increases the registration accuracy, which is also shown by Siebert et al. [7]. In image segmentation, weak label supervision has already gained interest. Rajchl et al. [8], for example, use an extension of the GrabCut algorithm and learn segmentation from bounding box annotations. In this paper, our aim is to close the gap between supervised and unsupervised registration methods and propose a weakly supervised learning scheme for deformable image registration including large deformations and introduce a loss function based on 3D bounding boxes to decrease the effort of the labeling process. We use inter-patient 3D Abdominal CT images and are able to increase the overlap of organs by approximately 10% in comparison to unsupervised image registration methods. If the time of the labeling process is taken into account, the performance of supervised algorithms can also be exceeded.

2 Methods

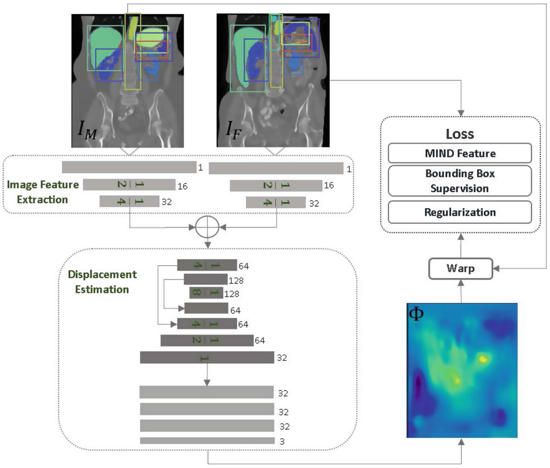


Fig. 1. Architecture of proposed method: Image features are extracted for I_F and I_M separately in two decoders (shared weights). The concatenated features are passed through a U-Net-like architecture and are finally used to estimate a displacement Φ to warp I_M. The loss consists of three parts: MIND features, regularization and the proposed bounding box supervision. The resolution in relation to the input resolution of the different steps are displayed in the layers.

The network consists of two parts: an image feature extraction part and a displacement estimation part. An overview of the architecture is shown in Fig. 1. The image feature extraction part extracts the low level features of the input images in two streams (with shared weights for monomodal registration). The displacement estimation part uses the concatenated low level features and estimates the displacement field. The 32 concatenated feature maps of I_F and I_M are used as input to extract 32 joint feature maps with a U-Net-like network with three encoder and four decoder blocks. Three additional sequences are added to estimate the displacement field. The final displacement field is generated by reducing the 32 feature maps to the three displacement dimensions with a 1 × 1 × 1 convolution and transformed to normalized sampling voxel locations (value range from -1 to 1) with the *tanh* activation function to match the PyTorch grid definition. The deformation has the same size as the input images.

To train the network, weak label supervision is used. Instead of using detailed labels for the calculation of the loss function, bounding boxes are used. The advantage of this method is that a significant reduction in time can be achieved and the variance between raters is also lower. A combination of three loss functions is used: the modality independent neighbourhood descriptor (MIND) with self-similar context (SSC) [10], a diffusion regularization and the mean squared error for the bounding boxes. The bounding box loss is multiplied by a factor of two.

To generate the final registration result including large deformations, we apply the network twice. The first input images are I_F and I_M. Then, I_M is warped with the first displacement field. The resulting warped moving image is used as second input.

3 Experiments

To train and evaluate our method, we use the publicly available Learn2Reg challenge dataset (Task3, 2020). This dataset contains 30 abdominal CT scans with thirteen manually labeled abdominal organs [4,5]. For training and testing, we use the split and validation pairs as in the official challenge. The data is already preprocessed to same voxel sizes and spatial dimensions. We downsample the images for the experiments to a size of 144 × 112 × 144 due to GPU memory requirements. For all labels, tight bounding boxes as well as a bounding box with a random error of ±5% are generated. The network is trained using Adam optimizer with a learning rate of 0.001 for 7500 iterations.

We train our network three times: unsupervised (not using the label loss), with the proposed bounding box loss, and with the voxelwise manually labeled organ segmentations. To establish comparability between training with label and weak label loss, we perform additional runs of supervised training with less training data. In this way, we simulate manual generation of labels or bounding boxes that takes the same amount of time. In total, we have five experiments: unsupervised, tight-weakly-supervised, weakly-supervised, supervised and supervised_50%. Tight-weakly refers to perfect bounding boxes, weakly refers to bounding boxes with an additional error of ±5% and supervised_50% refers to the experiment with less labeled data.

4 Results

In Table 1 the average Dice scores for all organs are listed for the different trainings. In comparison to the initial overlap of the organs, the overlap can

Table 1. Dice scores [%] for spleen 🟦, right kidney 🟦, left kidney 🟦, gall bladder 🟦, esophagus 🟩, liver 🟩, stomach 🟩, aorta 🟨, inferior vena cava 🟨, portal and splenic vein 🟧, pancreas 🟥, left adrenal gland 🟥, and right adrenal gland 🟫.

	🟦	🟦	🟦	🟦	🟩	🟩	🟩	🟨	🟨	🟧	🟥	🟥	🟫	avg ± std
initial	42	34	35	2	23	62	24	33	36	5	15	8	9	25 ± 13
unsupervised	67	57	61	5	33	81	35	54	50	15	21	18	14	39 ± 14
tight-weakly-supervised	70	67	69	7	33	86	41	53	56	20	27	25	17	44 ± 13
weakly-supervised	67	64	64	6	32	83	40	54	56	18	28	24	16	43 ± 13
supervised	81	73	78	8	43	86	50	67	61	17	25	21	16	48 ± 11
supervised-_50%	67	55	59	6	38	81	39	51	42	10	18	23	9	38 ± 13

be increased by approximately 14%. For the tight bounding box training, the overlap can be increased by approximately 19% and 18% for the bounding box training with random error. The label supervised trained network increased the overlap by approximately 22%. The standard deviation of the Jacobian determinant as well as the proportion of negative values are comparable for all trainings. It can be shown that a higher Dice score can be obtained for larger organs or for organs that initially already have a high overlap. The largest organ, the liver, for example, has the highest initial Dice overlap of 62%, and also the highest Dice overlap after registration for all variants (in a range of 81–85%). Organs with a small initial overlap, e.g. left adrenal gland (initial overlap 8%), also have a relatively low overlap after registration for all methods (in a range of 18–25%). For these organs, however, the Dice of weakly-supervised is higher than for supervised (e.g. left adrenal gland: 25% for weakly-supervised and 21% for supervised).

5 Discussion and Conclusion

We presented a deep-learning-based method for deformable image registration with weak bounding box supervision. We compared our method with an unsupervised and a label supervised training. The resulting registration of our method shows an improvement of about 5% for the Dice overlap in comparison to the unsupervised training. To simulate a realistic annotation of bounding boxes, we added an inter-observer-error of 5% per bounding box side, and showed that the quality of the result does not change significantly (approximately 1%) compared to tight bounding boxes. Organs with small initial overlap show the highest Dice score after the registration with the weak bounding box supervised network.

If the time for the labeling process was taken into account, so that less labels are available than bounding boxes, the accuracy of the label supervised training is less than for our bounding box supervision. Hence, for the purpose of medical image registration the proposed weak supervision strategy (labeling more images with lower effort) is beneficial.

References

1. de Vos, B.D., Berendsen, F.F., Viergever, M.A., Sokooti, H., Staring, M., Isgum, I.: A deep learning framework for unsupervised affine and deformable image registration. Med. Image Anal. **52**, 128–143 (2019)
2. Eppenhof, K.A., Pluim, J.P.: Pulmonary CT registration through supervised learning with convolutional neural networks. IEEE Trans. Med. Imag. **38**(5), 1097–1105 (2018)
3. Sentker, T., Madesta, F., Werner, R.: GDL-FIRE^{4D}: deep learning-based fast 4D CT image registration. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11070, pp. 765–773. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00928-1_86
4. Hansen, L., Hering, A., Heinrich, M.P., et al.: Learn2Reg: 2020 MICCAI registration challenge (2020). https://learn2reg.grand-challenge.org
5. Xu, Z., Lee, C.P., Heinrich, M.P., et al.: Evaluation of six registration methods for the human abdomen on clinically acquired CT. IEEE Trans. Biomed. Eng. **63**(8), 1563–1572 (2016)
6. 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
7. Siebert, H., Hansen, L., Heinrich, M.P.: Evaluating design choices for deep learning registration networks. In: Bildverarbeitung für die Medizin 2021. I, pp. 111–116. Springer, Wiesbaden (2021). https://doi.org/10.1007/978-3-658-33198-6_26
8. Rajchl, M., et al.: DeepCut: object segmentation from bounding box annotations using convolutional neural networks. IEEE Trans. Med. Imag. **36**(2), 674–683 (2016)
9. Hering, A., Kuckertz, S., Heldmann, S., Heinrich, M.P.: Memory-efficient 2.5 D convolutional transformer networks for multi-modal deformable registration with weak label supervision applied to whole-heart CT and MRI scans. Int. J. Comput. Assist. Radiol. Surg. **14**(11), 1901–1912 (2019)
10. Heinrich, M.P., Jenkinson, M., Papież, B.W., Brady, S.M., Schnabel, J.A.: Towards realtime multimodal fusion for image-guided interventions using self-similarities. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013. LNCS, vol. 8149, pp. 187–194. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40811-3_24

Author Index