Unsupervised Deformable Image Registration Using Polyphase UNet for 3D Brain MRI Volumes

Antoinette D. Martin, Boah Kim, Jong Chul Ye
Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea

Purpose: Image registration is a fundamental task in various medical imaging studies and clinical image analyses, such as comparison of patient data with anatomical structures. In order to solve the problems of conventional image registration approaches, such as long computational time, recent deep-learning supervised and unsupervised methods have been extensively studied because of their excellent performance and fast computational time. In this study, we propose a deep-learning-based network for deformable medical image registration using unsupervised learning.

Materials and Methods: In this paper, we solve the image-registration optimization problem by modelling a function using a convolutional neural network with polyphase decomposition to learn the spatial transformable parameters based on the input images and to generate the registration field. A spatial transformer is used to reconstruct the output warped image while imposing smoothness constraints on the registration field. With polyphase decomposition, our proposed method learns more features based on the input image pairs without the need for any ground-truth registration field.

Results: Experimental results using 3D T1 brain MRI volume scans and compared with state-of-the-art image-registration methods demonstrated that our method provides better 3D-image registration.

Conclusion: Our proposed method uses less computational time in registering unseen pairs of input images during inference and can be applied for other unimodal image registration tasks, and the hyper-parameters can be adjusted for the specific task.

Keywords: Polyphase decomposition; Unsupervised learning; Medical image registration

INTRODUCTION

Image registration is defined as aligning two (or more) given images based on identical geometrical coordination (1), which have matched imaging contents or similar structures. The images are acquired from different viewpoints, at different times, or using different sensors or modalities (2). The aim is to find an optimum spatial transformation that registers the structures of interest in the best way (1). It is used in many clinical applications, such as monitoring tumor growth, treatment verification, and comparison of patient data with anatomical structures. To solve the optimization
problem, conventional image-registration techniques have been proposed (1). These conventional methods involve an iterative process for a given pair of images. Thus, the process includes extracting proper features, selecting a similarity measure (to identify the amount of correspondence between the input images), estimating the transformer parameters or mapping function (i.e., choosing the model of transformation), applying the parameters to the moving image (which should lead to a new supposedly better-aligned image), and finally an evaluation measure to validate the process. If the termination criteria are satisfied, the algorithm is terminated and the output image is achieved, else a new iteration begins. The drawbacks of the conventional iterative method are its difficulty in being applied to most similarity measures, especially in multimodal image registration and not being applicable in real-time clinical applications because of the long computational time (1).

In order to solve the problem of the traditional similarity measures, deep-learning-based generalized similarity measures (3-6) have been proposed for better registration results. For example, Cheng et al. (7) trained a 5-layer stacked denoising autoencoder (SDAE) to approximate the similarity between a couple of input computed tomography (CT) and magnetic resonance (MR) images of the brain. It could also be referred to as a binary classifier. The binary output refers to the correspondence or non-correspondence between the image patches; the output before the sigmoid activation function is used to compute the similarity measure. The measure worked better than normalized mutual information (NMI) and local cross correlation (LCC).

The long computational time can be shortened by adopting supervised learning-based methods that have been proposed (8-12) to do registration in a single step instead of using the iterative process. This allows real-time, robust registration across various image domains. However, the quality of performance depends on the quality of the ground-truth registration fields or anatomical landmarks, which are difficult to obtain in medical imaging. Most of the ground-truth warp or deformation fields are obtained from classical registration methods on a pair of scans or simulated deformations and deformed images (13).

To solve the problem of supervision and provide neural networks with spatial transform capabilities, a spatial transformer module was proposed (14). The spatial-transformer module is a differentiable module, which applies a spatial transformation to a feature map during a single forward pass, where the transformation is conditioned on the particular input, producing a single-output feature map. It consists of a localization network that computes spatial transformation parameters for the input feature map, a grid generator that is intended to apply the estimated transformation parameters by the localization network to the input feature map, and a sampler that works as an interpolator to construct the final output warped image. It is computationally fast, so does not impair training speed. However, it is not specific for medical-imaging applications, because its deformation field does not preserve the nature of the anatomical structures in the body.

Inspired by Spatial Transformer Networks, several unsupervised methods have been proposed (15-20). An example is the VoxelMorph-2 method (21), whose aim is to speed up the computation for medical image analysis and processing tasks. Given two images, the convolutional neural network (CNN) models a function that learns the spatial parameters and generates the registration field and a spatial transformer function that warps the registration field with the moving image to construct the output image while imposing smoothness constraints on the registration field. Compared to other methods (15, 16), the method performs registration on full 3D MRI brain scans and handles the large deformations allowed by the registration field.

Although the VoxelMorph-2 framework works fairly well, it adopts the CNN architecture based on UNet (22); a much better outcome can be achieved by incorporating some recent advances in deep learning. Thus, we propose a CNN architecture that uses polyphase decomposition or lossless decomposition (23), which satisfies the frame condition in which low- and high-frequency components of images can be retained (24) to improve the performance of the VoxelMorph-2 framework. This allows more details from the input images to be captured and used in estimating or learning the spatial transformation parameters. The use of UNet with Polyphase decomposition, as reported in several recent research works, has improved the performance of some medical-image processing tasks (23, 25). In this study, we use an atlas-based registration, in which we compute a registration field between an atlas or reference volume and each volume in our dataset. We evaluate our method on 3D structural brain MR images from a variety of age groups using volume overlap of anatomical segmentations and compare our results with the VoxelMorph-2 architecture. The results show a higher performance with the proposed architecture.
MATERIALS AND METHODS

Let $F, M$ represent the two image volumes, where $F$ is the fixed image and $M$ the moving image. We use a CNN to module a function $g_\theta(F,M)=\phi$, where $F$ and $M$ are single-channel grayscale image volumes defined over a 3D spatial domain $\Omega \subset \mathbb{R}^3$; $\theta$ refers to the learnable parameters of $g$; and $\phi$ is the registration field. We then use a spatial transformer to warp $M(p)$ to $M(\phi(p))$, where $p$ refers to the voxel $p \in \Omega$ and $\phi(p)$ is a location such that $F(p)$ and $M(\phi(p))$ depict similar anatomical locations; we then evaluate the similarity of $M(\phi)$ and $F$ to update $\theta$. We used a stochastic gradient to find the optimal parameters $\theta$ by minimizing the expected loss function, using the training set. The overview of our method is presented in Figure 1.

Proposed CNN Architecture

UNet has been one of the main workhorses for medical image segmentation; however, it is limited, in that segmented boundaries tend to lose details. In order to solve this problem, we proposed Polyphase UNet, which incorporates polyphase decomposition or lossless decompositions that retains accurate boundary details, retains high-frequency components, and causes no change in polarity of the signal (23). Thus, instead of using max or average pooling, whose operation reports a single output within a rectangular neighbourhood, the $2^n$ neighbor pixels in the n-D input images are decomposed into $2^n$ channels of data with reduced size at the pooling layer, whereas in the un-pooling layer, the $2^n$ channels are grouped together into an enlarged single channel. Figure 2 shows the process for subpixel pooling and un-pooling for a 2D network. In our

Fig. 1. The overall framework of the proposed method for image registration.

Fig. 2. The process of subpixel pooling and un-pooling for a 2D network in which the channel is decomposed into 4 channels at pooling and combined into 1 at un-pooling.
network, since \( n = 3 \), the input is gathered into 8 channels for pooling and into 1 for un-pooling. This modification allows more details to be captured during the learning process. In comparison with UNet, it worked better in the liver and lesion segmentation task.

The proposed CNN architecture in this work is like that of the Polyphase U-Net (23) and the VoxelMorph-2 (21). The network consists of an encoder and decoder with skip connections to generate \( \phi \) given \( F \) and \( M \). The network takes a single input of the concatenation of \( F \) and \( M \) into a 2-channel 3D image. The input size is given as \( 160 \times 192 \times 224 \times 2 \). We apply 3D convolutions, Leaky ReLU activation, and batch normalization, with a convolutional kernel size of \( 3 \times 3 \times 3 \) in both the encoder and the decoder stages. The convolutional layers capture hierarchical features of the input image pair that are needed to compute \( \phi \). The encoding stage consists of convolutions (followed by Leaky ReLU activations), Batch Normalization, and Subpixel pooling. The pooling layer reduces the spatial dimensions by half until the smallest layer is achieved, which is \( 1/16 \) the size of the input images. In the decoding stage, we alternate between subpixel un-pooling, convolutions (followed by Leaky ReLU activations), Batch Normalization, and concatenating skip connections. With skipped connections, features learned from the encoding stages are propagated to layers in the decoder to aid with the registration. The output size of the decoder stage is \( 160 \times 192 \times 224 \times 3 \). The architecture is shown in Figure 3.

### Spatial Transformation Function

The spatial transformation layer proposed by [14] is added to the network to deform \( M \) with the registration field \( \phi \). This layer applies the transformation parameters from our network to \( M \) and then constructs the final warped image by interpolation. For our registration experiment, we used the spatial-transformation function with a trilinear interpolation to warp the image \( M \) with \( \phi \) to construct the output warped image \( M(\phi) \). The function is written as

\[
M(\phi(p)) = \sum_{q\in Z(\phi(p))} M(q) \prod_{d\in\{x,y,z\}} (1-|\phi_d(p)-q_d|)
\]

where \( Z(\phi(p)) \) are the 8-voxel cubic neighbors of \( \phi(p) \), and \( d \) refers to the directions in the 3D-image space. The differential ability of the operations allows backpropagation of errors during optimization.

### Loss Function

We train the network by solving the loss function:

\[
L = L_{sim} + \lambda L_{smooth}
\]

This function is based on the energy function of classical image-registration techniques. \( L_{sim} \) measures the image (dis) similarity between the output warped image and the fixed one, and \( L_{smooth} \) evaluates the smoothness of the registration field, \( \phi \). We use the negative local cross-correlation, which is robust to various intensities often found across scans, to penalize the differences or calculate the similarity between \( F \) and \( M(\phi) \). The cross-correlation is defined as

\[
CC(x,y) = \frac{|\langle x-x', y-y' \rangle|^2}{||x-x'|| ||y-y'||}
\]

**Fig. 3.** The diagram of the CNN architecture in the proposed method with the number of channels printed underneath each box, which represents a 3D volume.
where $\bar{x}$ and $\bar{y}$ denote the mean values of $x$ and $y$. A higher CC indicates better alignment, yielding $L_{\text{sim}} = -CC(F, M(\theta))$. $L_{\text{smooth}}$ is calculated by using the $l_2$-loss as a regularization function:

$$L_{\text{smooth}}(\theta) = \sum_{p \in \Omega} ||\nabla \theta(p)||^2$$

[4]

The complete loss function is given as

$$L = -CC(F, M(\theta)) + \lambda \sum_{p \in \Omega} ||\nabla \theta(p)||^2$$

[5]

where $\lambda$ is the regularization parameter.

Experiment

We evaluated the proposed method on the OASIS3 Dataset (26), which consists of 3D brain MRI scans acquired from 1,098 participants (including normal adults as well as individuals at various stages of cognitive decline) ranging from 42 to 95 years. The dataset consists of more than 2000 MR sessions, but in this experiment, we focused on the T1-weighted brain MRI scans. The OASIS3 dataset also provided FreeSurfer (27) files, which contain volumetric segmentations. We used the segmentation maps in evaluating our registration. With the provided FreeSurfer files, the total number of scans used was 2045, which we divided into 1636, 306, and 103 volumes for training, validation, and test sets, respectively.

The scans were resampled to a $256 \times 256 \times 256$ grid with 1-mm isotropic voxels using FreeSurfer. We then cropped the scans to $160 \times 192 \times 224$ and normalized them by dividing the volume/pixels by 255.0. We focused on atlas-based registration, in which the computed registration field is between an atlas or reference volume and each volume in the dataset. The atlas represents a reference (or average) volume and is usually constructed by jointly and repeatedly aligning a dataset of brain MR volumes and averaging them. The atlas file used is the same as the atlas file in VoxelMorph-2 (21). An input pair for the network consisted of the atlas (image F) and a random volume from the dataset (image M).

We implemented the network using Keras with a TensorFlow backend and used an Adam Optimizer with a learning rate of $1 \times 10^{-4}$. We used a batch size of one (1) in order to reduce memory usage. The number of epochs was 100, and the number of iterations equaled 160,000. We set the regularization parameter to 1.

We evaluated our method with the provided volumetric anatomical segmentations. With FreeSurfer, the brain can be segmented into more than 100 segments, but we focused on 30 structures. The deformed segmentation maps for the registered images were obtained by deforming the original segmentation maps of the moving images with the registration fields computed from the network. We measured the accuracy of our method by computing the Dice score between the deformed segmentation map and the atlas segmentation map. The Dice score, which quantifies the volume overlap between two structures, can be computed as:

$$\text{Dice}(A,B) = 2 \times \frac{A \cap B}{A + B}$$

[6]

where $A \cap B$ refers to the number of pixels that overlap between the structures $A$ and $B$, and $A + B$ refers to the total number of pixels in the structures $A$ and $B$. The Dice score ranges from zero (0) to one (1), with the value of 1 indicating greater similarity between the structures.

RESULTS

Figure 4 shows the average Dice scores for the test set for different values of the regularization parameter $\lambda$. The optimal Dice score occurs when $\lambda=1$. When there is no regularization ($\lambda=0$), the Dice score is higher than the Dice score of the affine registration. Figure 5 shows examples of the registration fields with different values of the regularization parameter. As the regularization parameter increases, the registration field becomes smoother. To evaluate the proposed method, we compared our method with ANTs (28), VoxelMorph-2, and a baseline of only affine registration. Table 1 shows the average dice score over all subjects in the test dataset for affine alignment, ANTs, VoxelMorph-2, and the proposed method, which worked slightly better than did the VoxelMorph-2 or ANTs, and all three show much improvement on the affine registration. We depict the distribution of Dice scores for each structure as boxplots in Figure 6. The proposed method achieves Dice measures comparable to those of ANTs and VoxelMorph-2 for all structures, worked better than VoxelMorph-2 on some structures, such as the hippocampi (both left and right) and third and fourth ventricles, and worse on others, such as the cerebral white matter (both left and right). Table 1 also shows the computation time for the ANTs, VoxelMorph-2, and the proposed method. From the table, both the VoxelMorph-2 and the proposed method compute a registration task within a few seconds when using
GPU. The VoxelMorph-2 results are achieved by using the existing model file provided on their GitHub page to deform the moving files and their respective segmentation files. Figures 7 and 8 show some examples of the MR coronal and axial slices with results from VoxelMorph-2 and the proposed method. The yellow and blue boxes highlight some anatomical structural differences between the results.

**DISCUSSION AND CONCLUSION**

In this study, we proposed a deformable registration network that incorporates polyphase decomposition in CNN to learn the features of the provided input volume pair and

| Method      | Average Dice | Time (min) |
|-------------|--------------|------------|
| Affine      | 0.621 (0.171)| 0          |
| ANTs SyN    | 0.752 (0.140)| 102 (CPU)  |
| VoxelMorph-2| 0.748 (0.143)| 0.01 (GPU) |
| Proposed    | 0.756 (0.137)| 0.01 (GPU) |

The standard deviations for the average dice scores are in parentheses.

---

**Table 1. Average Dice score over all structures and subjects in the test set and computational time for affine alignment, ANTs, VoxelMorph-2, and the proposed method**

Fig. 4. The effects of varying the regularization parameter $\lambda$ on the computed average Dice scores. The optimal score or best score was achieved when $\lambda=1$.

Fig. 5. Example of MR coronal slices of the proposed registration fields for different values of $\lambda$. Each row represents a different scan in the test set.
Fig. 6. Resulting boxplots of the Dice scores for the 30 anatomical structures for ANTs, VoxelMorph-2, and the proposed method.

Fig. 7. Example of MR coronal slices extracted from input pairs (columns 1-2) and the resulting warped slices for VoxelMorph-2 and the proposed method with highlighted differences in structures.
produce a registration or deformation field. Using polyphase decomposition, our proposed method is trained to find the mapping between the pair of input images without the use of any supervised information, such as landmarks or ground-truth registration fields. With the provided anatomical segmentations, we evaluated the accuracy of the method using the Dice Score as a proxy measure of registration accuracy. Though there is a slight difference in the Dice scores of the proposed method from those of VoxelMorph-2 (used as the benchmark), this proves that with polyphase decomposition instead of max-pooling or average pooling, CNNs tend to learn more features about the input image pairs. For conventional registration methods, which produce an optimization function for each test image pair, our method uses an optimization function that is aggregated over a dataset during the training phase. Our method can be applied for other unimodal image-registration tasks, and the hyper-parameters can be adjusted for the specific task.

REFERENCES

1. Boveiri HR, Khayami R, Javidan R, MehdiZadeh AR. Medical image registration using deep neural networks: a comprehensive review. ArXiv 2002.03401;1-45
2. Haskins G, Kruger U, Yan P. Deep learning in medical image registration: a survey. Mach Vision Appl 2020;31:8
3. Wu G, Kim M, Wang Q, Gao Y, Liao S, Shen D. Unsupervised deep feature learning for deformable registration of MR brain images. Med Image Comput Comput Assist Interv 2013;16:649-656
4. Ghosal S, Ray N. Deep deformable registration: enhancing accuracy by fully convolutional neural net. Pattern Recognit Lett 2017;94:81-86
5. Blendowski M, Heinrich MP. 3D-CNNs for deep binary descriptor learning in medical volume data. Informatik Aktuell 2018:23-28
6. Liu X, Jiang D, Wang M, Song Z. Image synthesis-based multi-modal image registration framework by using

Fig. 8. Example of MR axial slices extracted from input pairs (columns 1-2) and resulting warped slices for VoxelMorph-2 and the proposed method with highlighted differences in structures.
deep fully convolutional networks. Med Biol Eng Comput 2019;57:1037-1048
7. Cheng X, Wang B, Liu Y, Yuan Y, Shu Y, Chen B. Comparable electrode impedance and speech perception at 12 months after cochlear implantation using round window versus cochleostomy: an analysis of 40 patients. ORL J Otorhinolaryngol Relat Spec 2018;80:248-258
8. Yang X, Kwitt R, Niethammer M. Fast predictive image registration. Lect Notes Comput Sc 2016:48-57
9. Sokooti H, de Vos B, Berendsen F, Lelieveldt BPF, Işgum I, Staring M. Nonrigid image registration using multi-scale 3D convolutional neural networks. Lect Notes Comput Sc 2017:232-239
10. Eppenhof KAJ, Lagarge MW, Moeskops P, Veta M, Pluim JPW. Deformable image registration using convolutional neural networks. Med Imaging 2018: Image Processing 2018:105740S
11. Cao X, Yang J, Zhang J, et al. Deformable image registration based on similarity-steered CNN regression. Med Image Comput Assist Interv 2017;10433:300-308
12. Sun L, Zhang S. Deformable MRI-ultrasound registration using 3D convolutional neural network. Lect Notes Comput Sc 2018:152-158
13. Balakrishnan G, Zhao A, Sabuncu MR, Guttag J, Dalca AV. VoxelMorph: a learning framework for deformable medical image registration. IEEE Trans Med Imaging 2019 [Online ahead of print]
14. Jaderberg M, Simonyan K, Zisserman A, Kavukcuoglu K. Spatial transformer networks. Advances in Neural Information Processing Systems, 2015:2017-2025
15. Li H, Fan Y. Non-rigid image registration using self-supervised fully convolutional networks without training data. IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), 2018:1075-1078
16. de Vos BD, Berendsen FF, Viergever MA, Staring M, Işgum I. End-to-end unsupervised deformable image registration with a convolutional neural network. Lect Notes Comput Sc, 2017:204-212
17. Dalca AV, Balakrishnan G, Guttag J, Sabuncu MR. Unsupervised learning for fast probabilistic diffeomorphic registration. Lect Notes Comput Sc, 2018:729-738
18. Yoo I, Hildebrand DGC, Tobin WF, Lee WCA, Jeong WK. ssEMnet: serial-section electron microscopy image registration using a spatial transformer network with learned features. Lect Notes Comput Sc 2017:249-257
19. Shu C, Chen X, Xie Q, Han H. An unsupervised network for fast microscopic image registration. Medical Imaging 2018: Digital Pathology 2018:105811D
20. Kearney V, Haaf S, Sudhyadhom A, Valdes G, Solberg TD. An unsupervised convolutional neural network-based algorithm for deformable image registration. Phys Med Biol 2018;63:185017
21. Balakrishnan G, Zhao A, Sabuncu MR, Guttag J, Dalca AV. An unsupervised learning model for deformable medical image registration. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018:9252-9260
22. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. MICCAI 2015: Medical Image Computing and Computer-Assisted Intervention, 2015:234-241