# Deformable Image Registration Based on Similarity-Steered CNN Regression

Xiaohuan Cao[1,2], Jianhua Yang[1], Jun Zhang[2], Dong Nie[2], Minjeong Kim[2], Qian Wang[3], and Dinggang Shen[2(✉)]

[1] School of Automation, Northwestern Polytechnical University, Xi'an, China
[2] Department of Radiology and BRIC, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
dgshen@med.unc.edu
[3] School of Biomedical Engineering, Med-X Research Institute, Shanghai Jiao Tong University, Shanghai, China

**Abstract.** Existing deformable registration methods require exhaustively iterative optimization, along with careful parameter tuning, to estimate the deformation field between images. Although some learning-based methods have been proposed for initiating deformation estimation, they are often template-specific and not flexible in practical use. In this paper, we propose a convolutional neural network (CNN) based regression model to directly learn the complex mapping from the input image pair (i.e., a pair of template and subject) to their corresponding deformation field. Specifically, our CNN architecture is designed in a patch-based manner to learn the complex mapping from the input patch pairs to their respective deformation field. First, the equalized active-points guided sampling strategy is introduced to facilitate accurate CNN model learning upon a limited image dataset. Then, the similarity-steered CNN architecture is designed, where we propose to add the auxiliary contextual cue, i.e., the similarity between input patches, to more directly guide the learning process. Experiments on different brain image datasets demonstrate promising registration performance based on our CNN model. Furthermore, it is found that the trained CNN model from one dataset can be successfully transferred to another dataset, although brain appearances across datasets are quite variable.

## 1 Introduction

Deformable registration is a fundamental image processing step for many medical image analysis tasks since it can help build anatomical correspondences across images. Among existing deformable registration algorithms, most of them regard image registration as a high-dimensional optimization problem, which intends to maximize the similarity between the template and subject images with a regularization upon the deformation field. Generally, these methods often require iterative optimization to estimate the deformation field between images, as well as careful parameter tuning. Moreover, the registration performance may decline significantly when existing large appearance variation between the template and the *to-be-registered* subject.


M. Descoteaux et al. (Eds.): MICCAI 2017, Part I, LNCS 10433, pp. 300–308, 2017.
DOI: 10.1007/978-3-319-66182-7_35

Some learning-based methods [1–4] are proposed to predict the initial deformation field or parameters for registration. Then, the roughly predicted deformation field can be refined by adopting existing registration algorithms in an effective manner. Although these methods can partially improve the performance of registration, there are still some limitations. (1) The learning is *template-specific*, while changing the template requires re-training from scratch. (2) The prediction models often ignore the intrinsic matching associations between the *to-be-registered* image pair along with their local correspondence. (3) The predicted deformation field still needs further refinement, i.e., by employing a conventional registration method.

Recently, deep learning techniques such as convolutional neural network (CNN) become well known for their strong end-to-end learning ability. In this paper, we propose to learn a general CNN-based regression model, in order to directly construct a mapping from the input image pair (e.g., a pair of template and subject) to their final deformation field. Then, in the application stage, we can input an unseen image pair to the CNN and effectively obtain the accurate deformation field between them. Our main contributions and novelties can be summarized as follows.

(1) To learn a general CNN regression model that is independent of any arbitrary template, we propose to regress from any image pair to their corresponding deformations. In particular, given two patches at the same locations of two different images, the CNN produces the displacement vector to align the two patches. A whole-image deformation field can then be derived accordingly, which relies on robust machine learning, rather than tedious parameter tuning in optimization.
(2) In order to bridge the large appearance gap between the pair of template and subject, we introduce an *auxiliary contextual cue* to guide the learning of the CNN. This cue encodes the easy-to-compute image patch similarities in a multi-scale way, which is shown to be important for successfully establishing the final deformation field and is also robust to large appearance variations.
(3) To make the CNN regression model more accurate, we introduce the *equalized active-points guided sampling strategy*, such that the training set complies well with the distributions of image patches and displacements. This strategy significantly enhances the accuracy when estimating the deformation field, and helps avoid further refinement by conventional registration methods.

## 2 Method

In this paper, we propose a similarity-steered CNN regression architecture to learn the mapping $\mathcal{M}$ from the image pair (e.g., a template $\mathcal{T}$ and a subject $\mathcal{S}$) to their final deformation field $\phi : \mathcal{M} : \ (\mathcal{T}, \mathcal{S}) \Rightarrow \phi$. Particularly, the inputs consist of two independent images. Obviously, our learning target is the local matching correspondence between the two input images. This is substantially different from the conventional CNN-based tasks.

As shown in Fig. 1, our CNN model is designed in a patch-wise manner to encode both the patch appearance information and the local displacement. **First**, two patches

are extracted from the same locations in template and subject. **Then**, we generate a multi-scale contextual cue to describe the similarity within the patch pair (*Network Part I*). The patches and the cue are concatenated as the multi-channel input to CNN, which regresses the final displacement vector for template patch center (*Network part II*). **Finally**, we predict the displacements for many sampled locations, and obtain the dense deformation field by thin-plate spline (TPS) interpolation in an iterative manner.

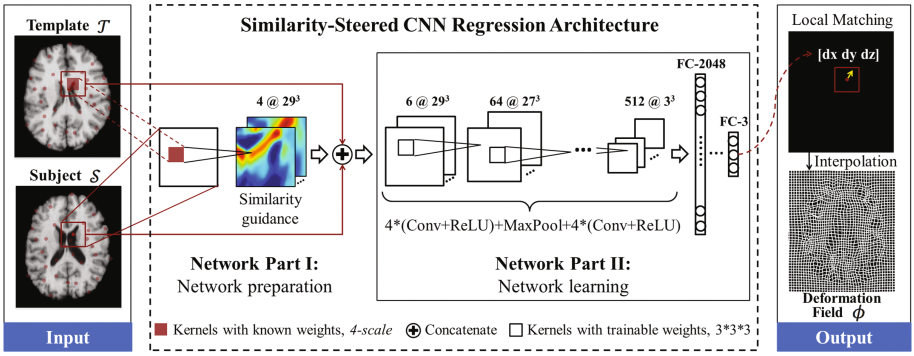


**Fig. 1.** The framework of the proposed similarity-steered CNN regression for deformable image registration. The input image pair has already been linearly aligned. Here, we use 2D examples for easy illustration, while our implementations are in 3D.

## 2.1 Training Set Preparation

For a pair of registered template image $\mathcal{T}$ and subject image $\mathcal{S}$ along with their deformation field $\phi$, a local patch pair $(p_{\mathcal{T}}(u), p_{\mathcal{S}}(u))$ is extracted from the center location $u$. We then obtain a training sample $S_i = \{(p_{\mathcal{T}}(u), p_{\mathcal{S}}(u))|\phi(u)\}$, where $\phi(u) = [d_x, d_y, d_z]$ is the displacement vector of $u$.

Obviously, a well-prepared training set is important to the accuracy and the robustness of the learned CNN. Conventional sampling often collects training patches randomly or uniformly in the input image spaces, while ignores the distribution of the displacements in the output space. Figure 2(a) presents the distribution of the displacement magnitudes measured from 20 real deformation fields (excluding background voxels). If the training patches are extracted randomly from the input image space only, the displacement magnitudes for >74% patches are below 1 mm. In this way, the generalization performance of CNN will be confined, which leads to underestimation of the displacement magnitude. An instance is shown in Fig. 2(b) and (c) for comparison. Therefore, we argue that all training patches should be sampled by referring to *not only* the input image space, *but also* the output displacement space.

In the **input *image* space**, we apply the *active-points guided sampling strategy*, where the importance $I(u)$ of each point $u$ can be related to gradient magnitude in the template image space. The voxel with rich anatomical information (e.g., strong edges) will have high importance to be sampled. Obviously, the density of the active-points will be higher on informative brain regions while lower on smooth regions.

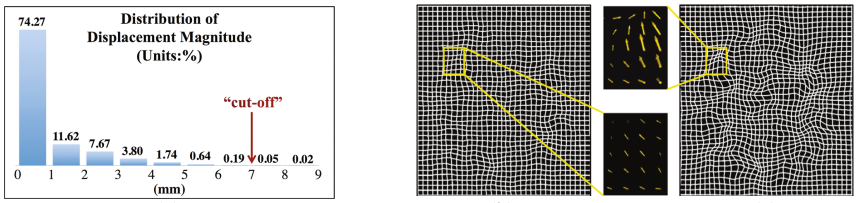


**Fig. 2.** Illustration of the *equalized sampling strategy*. The displacement magnitudes distribute unevenly as shown in (a) The deformation field is thus underestimated in (b) if following the conventional sampling strategy, and much improved in (c) when using our proposed strategy.

In the **output *displacement* space**, we adopt the *equalized sampling strategy* based on the displacement distribution. By incorporating the information from the input image space, we can sample the point $u$ with the integrated probability $P(u)$:

$$P(u) = e^{-\omega \cdot \left(\frac{\|\phi(u)\|_2}{\tau} I(u)\right)^{-1}}, \|\phi(u)\|_2 = \begin{cases} \left\|\phi(u)_2\right\|, & \|\phi(u)\|_2 \le \tau \\ 0 & \|\phi(u)\|_2 > \tau \end{cases} \tag{1}$$

Here, $\omega$ is a parameter to control the sampling probability as well as the sample number, and $\tau$ is a cut-off threshold. Apparently, the point $u$ with larger displacement magnitude $\|\phi(u)\|_2$ and importance $I(u)$ can be more likely to be sampled. However, the very large displacement is unpredictable concerning the limited modeling capability of CNN and the number of training patches. Thus, we apply the cut-off $\tau$ to saturate all displacements over the threshold.

After the *equalized active-points guided sampling*, the distribution of whole training set $S$ is mostly uniform within $U(0, \tau]$. In this paper, we set $\tau = 7$ mm. It is worth noting that, the displacement in the final deformation field is not limited by $\tau$. We iteratively perform the learned CNN model, such that the estimated displacements are accumulated to approximate the final deformation field.

### 2.2 Similarity-Steered CNN Regression

To bridge the large gap between the input image pair and the output displacement, we introduce the auxiliary contextual to guide CNN training. As shown in Fig. 1, our CNN model consists of two parts: (1) *network preparation* and (2) *network learning*.

**Network preparation.** The contextual cue is provided by the similarity map, which is the local cross-correlation from the center location in template patch to the whole subject patch locations, and we use a small image patch to represent each location, as shown in Fig. 3. In our implementation, we conduct it as a convolutional layer incorporated to the whole CNN architecture, in order to effectively obtain the similarity map $H$:

$$H = \frac{k_{\mathcal{T}}^{'}(u) * p_{\mathcal{S}}(u)}{|k_{\mathcal{T}}(u)| \cdot \|p_{\mathcal{S}}(u)\|}, \|p_{\mathcal{S}}(u)\| = \sqrt{k_1 * (p_{\mathcal{S}}(u) \cdot p_{\mathcal{S}}(u))} \tag{2}$$

where "$*$" is the convolution operation, and $k'_{\mathcal{T}}(u)$ is the reversed kernel derived from template patch $p_{\mathcal{T}}(u)$ at the center voxel $u$. For each patch pair, $k_{\mathcal{T}}(u)$ is fixed, thus the $L_2$-norm $|k_{\mathcal{T}}(u)|$ is a constant. $\|p_{\mathcal{S}}(u)\|$ is the $L_2$-norm map with the same size as subject patch $p_{\mathcal{S}}(u)$, where we also fast generate it by another convolution operation, i.e., convolving the self dot product $(p_{\mathcal{S}}(u) \cdot p_{\mathcal{S}}(u))$ with the kernel $k_1$. Here, $k_1$ is a kernel with all 1 elements and has the same size as $k_{\mathcal{T}}(u)$. Equation (2) can be identified as normalized cross-correlation. It is worth noting that, the kernel in this convolutional layer is derived from the data, so that, the weights are fixed and not trainable.

The similarity map allows us to establish correspondences between the two patches. However, the choice of the kernel affects the distinctiveness of correspondence, as an example shown in Fig. 3. Thus, we provide multi-scale similarity cue, corresponding to different sizes of kernels, to guide the training of CNN. In this paper, we use 4 kernel sizes as also illustrated in Fig. 3.

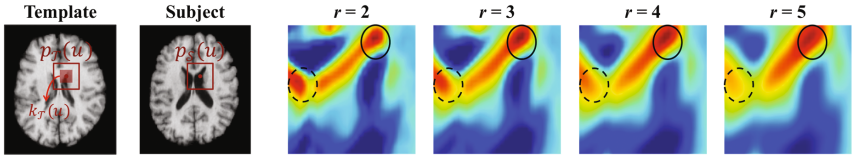


**Fig. 3.** Similarity maps of one sample with different kernel size $2r + 1$. Solid and dashed circle indicate correct guidance and incorrect guidance, respectively.

**Network learning.** The CNN architecture estimates the final displacement vectors with multi-channel inputs, including the patch pair and the similarity maps. Specifically, each convolution layer is followed by ReLU activations. The kernel number is doubled every two convolution layers, which starts from 64 to final 512 with the fixed size $3 \times 3 \times 3$. The subsequent fully connected (FC) layers consist of 3 layers with ReLU activations, and tanh activation for the final FC layer. The loss function is the mean squared error. It is worth noting that, padding operation is not applied in each convolution layer in order to avoid introducing meaningless information. The patch size will gradually decrease, and all neighborhood information of each sample point can be effectively incorporated in high dimensional space to help better represent the samples. Furthermore, only one pooling layer is adopted in order to protect the continuity of the regression model as well as make the network training efficiently.

## 3 Experiments

Two datasets, i.e. LONI LPBA40 and ADNI (Alzheimer's Disease Neuroimaging Initiative), are used to evaluate our registration performance. The LONI dataset contains 40 young adult brain MR images with 54 ROI labels, and additional tissue segmentations of gray matter (GM), white matter (WM) and cerebrospinal fluid (CSF). For ADNI, 30 brain MR images are randomly selected from the dataset, each of which has GM, WM and CSF segmentations. After preprocessing and affine registration, all

the images of the two datasets are resampled to the same size ($220 \times 220 \times 184$) and same resolution ($1 \times 1 \times 1$ mm$^3$).

The **training image data** is derived from LONI LPBA40. Specifically, we select 30 images to train, and test upon the remaining 10 images. We further randomly draw 30 image pairs from the training images. In order to obtain the very accurate deformation field of each image pair, SyN [5] is first applied on the intensity images with careful parameter tuning, and then Diffeomorphic Demons [6] is adopted on their tissue segmentation to further refine the registration accuracy. The final deformation field composed by this two-stage registration is used as ground-truth for CNN training. Then, 24K training samples are extracted from each image pair via the equalized active-points guided sampling strategy. In all, we have 720K training samples.

We train our similarity-steered CNN model on an Nvidia GPU by our modified 3D version of Caffe [7]. We start with the learning rate $\lambda = 0.01$ and multiply it by 0.5 after every 70K iterations. 20K samples are taken from the whole training set and used as validation data to monitor the overfitting problem. After training, we test the CNN model on the remaining 10 images in LONI LPBA40 and 30 images in ADNI. For each to-be-registered image pair, we estimate the displacements on 0.9% of all voxels that are selected by the *active-points guided sampling strategy*. The dense deformation field can be obtained by TPS interpolation [8]. We perform the above procedure for two iterations, and the incremental displacements are composed for the estimation of the final deformation field. Two popular state-of-the-art registration methods, i.e., SyN [5] and Demons [6], are chosen for comparison.

### 3.1 LONI Dataset

For the 10 testing subjects in LPBA40 dataset, we perform deformable image registration on each two images and report the averaged results in Fig. 4 and Table 1. Figure 4 shows the Dice similarity coefficient (DSC) on 54 brain ROIs. We observe

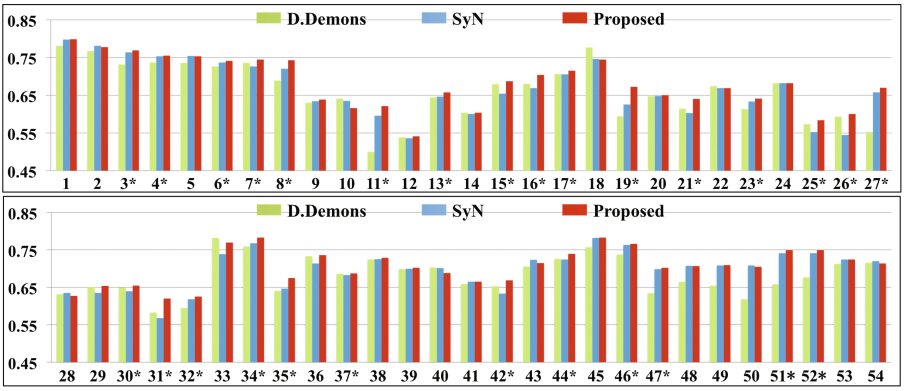


**Fig. 4.** Mean DSC of each of 54 ROIs based on 10 testing subjects from LONI dataset, after deformable registration by Demons, SyN, and our proposed method. "*" indicates statistically significant improvement by our proposed method, compared with other two methods ($p < 0.05$).

**Table 1.** Mean DSC and ASSD of GM, WM and CSF based on 10 testing subject from LONI dataset, after deformable registration by Demons, SyN, and our proposed method. "*" indicates statistically significant improvement by our method, compared to other two methods ($p < 0.05$).

| | DSC (%) | | | ASSD (mm) | | |
|---|---|---|---|---|---|---|
| | GM | WM | CSF | GM | WM | CSF |
| Demons | 72.8 ± 1.0 | 80.8 ± 7.3 | **62.9 ± 5.1** | 0.55 ± 0.34 | **0.49 ± 0.15** | **0.43 ± 0.10** |
| SyN | 72.7 ± 1.6 | 78.1 ± 0.7 | 61.5 ± 2.6 | 0.46 ± 0.03 | 0.58 ± 0.05 | 0.54 ± 0.05 |
| Proposed | **75.3 ± 1.4*** | **81.3 ± 0.6*** | 61.2 ± 2.4 | **0.43 ± 0.04*** | 0.56 ± 0.05 | 0.48 ± 0.03 |

that our method has better performances on 36/54 ROIs. Among them, 28 ROIs are statistical-significantly improved ($p < 0.05$) regarding both Demons and SyN.

Table 1 provides the DSC on the labels of GM, WM and CSF. Our method achieves significant improvements on GM and WM. In term of symmetric average surface distance (SASD) [9], we also obtain better performance on GM. Although the averaged accuracies of the competing methods are slightly higher than the proposed method in some regions, the differences, however, are not significant in paired *t*-tests.

This means that, we have at least achieved the comparable performance with the state-of-the-art deformable registration methods. Note that, our method only uses 0.9% test points to generate the whole deformation field, which leads to the reported performance, without exhaustive iterative optimization and parameter tuning. It suggests that the complex mapping from the image pair to the deformation field is successfully modeled by our proposed method.

### 3.2 ADNI Dataset

To further evaluate the transferring capability of the learned CNN, we test 30 ADNI images by directly applying the model trained on the LONI dataset. To enlarge the appearance variation between the to-be-registered image pair, in this experiment, 3 images are randomly selected from LONI dataset and used as templates. All 30 ANDI subjects are registered to those 3 templates, respectively, with results reported below.

Since only GM, WM and CSF labels are available for both these two datasets, we evaluate the registration performance based on these tissue labels in Table 2, and provide qualitative comparisons in Fig. 5. We observe that our proposed method achieves the best overall performance for this challenging registration task, with statistically significant improvements. Note that, even the image pair has large appearance

**Table 2.** Mean DSC and ASSD of GM, WM and CSF for ADNI dataset, after deformable registration by Demons, SyN, and our proposed method. "*" indicates statistically significant improvement, compared to other two methods ($p < 0.05$).

| | DSC (%) | | | ASSD (mm) | | |
|---|---|---|---|---|---|---|
| | GM | WM | CSF | GM | WM | CSF |
| Demons | 64.8 ± 2.3 | 75.7 ± 1.0 | 54.3 ± 2.5 | 0.70 ± 0.04 | **0.73 ± 0.05** | 0.38 ± 0.04 |
| SyN | 64.6 ± 2.5 | 76.3 ± 1.8 | 55.2 ± 3.2 | 0.69 ± 0.04 | 0.81 ± 0.06 | 0.37 ± 0.05 |
| Proposed | **65.1 ± 2.2*** | **78.3 ± 0.6*** | **56.0 ± 2.1*** | **0.64 ± 0.04*** | 0.75 ± 0.04 | **0.32 ± 0.04*** |

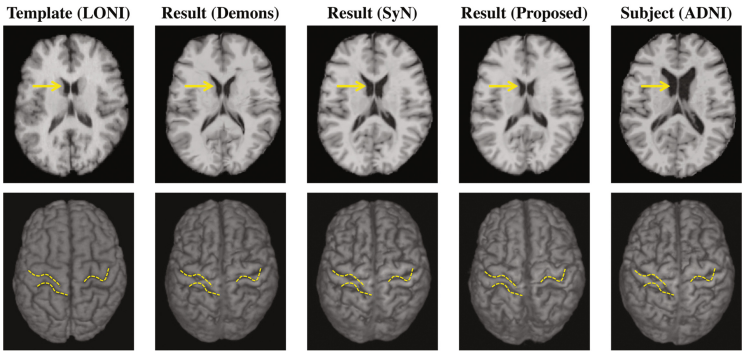


**Fig. 5.** Visualized registration results by Demons, SyN and our proposed method. Obvious improvements by the proposed method on ventricle region (first row), and the central sulcus and postcentral gyrus in 3D rending view (second row).

variation, our proposed method still obtains high performance without any parameter tuning. This indicates that the established CNN model is robust and accurate for complicated registration cases, which makes our method more flexible and applicable.

## 4 Conclusion

In this paper, we have proposed a novel deformable registration method by directly learning the complex mapping from the input image pair to the final deformation field via CNN regression. The equalized active-points guided sampling strategy is proposed, which facilitates training the regression model even with small dataset. Then, a similarity-steered CNN architecture is designed, where an additional convolutional layer is established in the whole network to provide similarity guidance during model learning. Experimental results show promising registration performance compared with the state-of-the-art methods on different datasets.

## References

1. Wang, Q., et al.: Predict brain MR image registration via sparse learning of appearance and transformation. Med. Image Anal. **20**(1), 61–75 (2015)
2. Yang, X., Kwitt, R., Niethammer, M.: Fast predictive image registration. In: Carneiro, G., et al. (eds.) LABELS/DLMIA -2016. LNCS, vol. 10008, pp. 48–57. Springer, Cham (2016). doi:10.1007/978-3-319-46976-8_6
3. Kim, M., et al.: A general fast registration framework by learning deformation–appearance correlation. IEEE Trans. Image Process. **21**(4), 1823–1833 (2012)
4. Gutiérrez-Becker, B., Mateus, D., Peter, L., Navab, N.: Learning optimization updates for multimodal registration. In: Ourselin, S., Joskowicz, L., Sabuncu, Mert R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9902, pp. 19–27. Springer, Cham (2016). doi:10.1007/978-3-319-46726-9_3

5. Avants, B.B., et al.: Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Med. Image Anal. **12** (1), 26–41 (2008)
6. Vercauteren, T., et al.: Diffeomorphic demons: efficient non-parametric image registration. NeuroImage **45**(1), S61–S72 (2009)
7. Jia, Y., et al.: Caffe: convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM International Conference on Multimedia. ACM (2014)
8. Zhang, J., et al.: Alzheimer's disease diagnosis using landmark-based features from longitudinal structural MR images. IEEE J. Biomed. Health Inform. (2017). doi:10.1109/JBHI.2017.2704614
9. Cao, X., et al.: Dual-core steered non-rigid registration for multi-modal images via bi-directional image synthesis. Med. Image Anal. **41**, 18–31 (2017)