Generating Patient-like Phantoms Using Fully Unsupervised Deformable Image Registration with Convolutional Neural Networks

Junyu Chen, Ye Li, Eric C. Frey
Russell H. Morgan Department of Radiology and Radiological Science, Johns Hopkins Medical Institutes, Baltimore, MD, USA
Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD, USA

The use of Convolutional neural networks (ConvNets) in medical imaging research has become widespread in recent years. However, a major drawback of these methods is that they require a large number of annotated training images. Data augmentation has been proposed to alleviate this. One data augmentation strategy is to apply random deformation to existing image data, but the deformed images often will not follow exhibit realistic shape or intensity patterns. In this paper, we present a novel, ConvNet based image registration method for creating patient-like digital phantoms from the existing computerized phantoms. Unlike existing learning-based registration techniques, for which the performance predominantly depends on the domain-specific training images, the proposed method is fully unsupervised, meaning that it optimizes an objective function independently of training data for a given image pair. While classical methods registration also do not require training data, they work in a lower-dimensional parameter space; the proposed approach operates directly in the high-dimensional parameter space without any training beforehand. In this paper we show that the resulting deformed phantom competently matches the anatomy model of a real human while providing the "gold-standard" for the anatomies. Compared with simulation programs, the generated phantoms could potentially serve as a data augmentation tool in today’s deep learning studies.

Index Terms—Image Registration, Computerized Phantom, CT, Convolutional Neural Networks

I. INTRODUCTION

COMPUTERIZED phantoms for nuclear medicine imaging research have been built based on anatomical and physiological models of human beings. They have played a crucial part in evaluation and optimization of medical imaging techniques and image reconstruction, processing and analysis methods [1]–[4]. Since the exact structural and physiological properties of the phantom are known, they can serve as a gold standard for the evaluation and optimization process. The 4D extended cardiac-torso (XCAT) phantom [5] was developed based on anatomical images from the Visible Human Project data. This realistic phantom includes parameterized models for anatomy, which allows the generation of different anatomical variations of the phantom. These phantoms have been used in Nuclear Medicine imaging researches [6]–[8], as well as in the various applications of deep learning [9]–[11]. By changing the values of parameters that control organ anatomy, the volumes and shapes of some tissues can be varied. However, the scaling of organs, even when different factors are used in orthogonal directions, does not fully and truly capture the interior anatomical variations within different patients. In [12], Segars et al. used a deformable image registration technique to map phantom labels to patient segmentation; the resulting deformation fields were then being applied to the phantom, thus creating a population of the new XCAT models that capture the anatomical variability among patients. This method relies on the segmentation of patient images, which is tedious and time consuming. In this work, we propose a Convolutional Neural Networks (ConvNets) based approach to perform patient to patient registration. The resulting deformation field can then be applied to organ label maps to automatically generate a segmentation of the patient image.

Deformable Image registration is a process of transforming two images into a single coordinate system, where one image is often referred to as the moving image, we denote it as $I_m$, while the other is referred to as the fixed image, denoted as $I_f$. Traditional methods formulate registration as a variational problem for estimating a smooth mapping, $\phi$, between the points in one image and those in another. They often tend to iteratively minimize the following energy function (eq. 1) on a single image pair [13]:

$$E = E_{sim}(I_m \circ \phi, I_f) + R(\phi),$$

where, $E_{sim}$ measures the level of alignment between the transformed moving image, $I_m \circ \phi$, and the fixed image, $I_f$. Some common choices for $E_{sim}$ are mean squared error (MSE) or the $L^2$ norm of the difference [14], sum of squared differences (SSD) [15], cross-correlation (CC) [16], and mutual information (MI) [17]. The transformation, $\phi$, at every point is defined by an identity transformation with the displacement field $u$, or $\phi = Id + u$, where $Id$ represents the identity transform [18]. The second term, $R(\phi)$, is referred to as the regularization on the deformation, $\phi$, which enforces the spatial smoothness. It is usually characterized by the gradients of $u$. One common assumption is that similar structures are presented in both moving and fixed images. Hence, a continuous and invertible deformation field (a diffeomorphism) is more desired, and the regularization term, $R(\phi)$, was designed for such reason. While diffeomorphisms are essential in some studies, for which the registration field is analyzed further. In the application of registration-based segmentation, the quality of the segmentation propagation is more critical than the diffeomorphic property of the underlying deformation fields [19]. In this study, due to the large interior and exterior shape
variability between digital phantoms and patients, we did not impose the registration to be diffeomorphic.

Recently, many deep learning-based methods were proposed to perform the registration tasks, for instance, [18], [20]–[23]. Some of the listed methods were introduced as the unsupervised (or more precisely, self-supervised) techniques, but they still require a prior training stage with a large amount of training data. These methods assume that neural networks could learn the universal computation of the displacement field by minimizing the registration energy function over a dataset of images. This is a common assumption to make with the deep learning based approach. Yet, such an assumption could be unreliable according to a recent study from Zhang et al. [24], where they showed that a well-generalized CNN classifier trained by a large dataset can still easily overfit a random labeling of the training data. More studies on fooling the deep neural networks (DNNs) with adversarial images also suggested that the well-trained networks can be unstable to small or even tiny perturbations of the data [25]–[29]. Whereas, our proposed registration method is fully unsupervised, meaning that no previous training is required. Whereas in image registration, instead of starting from a random initialization (i.e., random noises), it makes logical sense to initialize the ConvNet with moving images. Since one would like to make a moving image as similar to a target image as possible, an early stopping is not desired. In this work, we treat ConvNet as an optimization tool, where it minimizes the energy function via reparametrization in each iteration.

II. Method

A. Computerized Phantom Generation

The phantom used in this study was created on the 3D attenuation distributions of the realistic NURBS-based XCAT phantom [31]. Attenuation values were computed based on the material compositions of the materials and the attenuation coefficients of the constituents at 140 keV, the photon energy of Tc-99m used in Nuclear Medicine. Only a single 3D phantom image was used to be deformed to multiple patient CT image. Simulated attenuation map image can be treated as the template image, and phantom image can then be think of as the atlas in the traditional paradigm of medical image registration. Our aim is to register phantom image to patient CT images for the segmentation of patient scans and creating patient-like phantoms.

B. Image Registration with ConvNet

Let a moving image be $I_m$, and a fixed image be $I_f$, we assume that they are 2D grayscale images and affinely aligned. We model the computation of the displacement field, $\phi$, given the image pair, $I_m$ and $I_f$, using a deep ConvNet with its parameters $\theta$, i.e., $f_\theta(I_m, I_f) = \phi$. Figure 1 describes the architecture of our proposed method, it consists of a ConvNet that outputs registration field, and a B-spline spatial
First, the ConvNet generates the \( \phi \) for the given image pair of \( I_m \) and \( I_f \). Second, the deformed moving image is obtained by applying a B-spline spatial transformer that warps \( I_m \) with \( \phi \) (i.e., \( I_m \circ \phi \)). Finally, We backpropagate the loss from the similarity measure between \( I_m \circ \phi \) and \( I_f \) to update \( \theta \) in the ConvNet. The steps were repeated iteratively until the loss converges, then the resulting \( \phi \) represents the optimal registration field for the given image pair. The loss function \( (\mathcal{L}) \) of this problem can be formulated mathematically as:

\[
\mathcal{L}(I_m, I_f, \phi; \theta) = \mathcal{L}_{sim}(I_m \circ \phi, I_f; \theta) + \lambda R(\phi; \theta)
\]

Then, the parameters \( \theta \) that generate the optimal registration field can be enumerated by the minimizer:

\[
\theta^* = \arg \min_\theta \mathcal{L}(I_m, I_f, \phi; \theta),
\]

and the optimal \( \phi \) is given by:

\[
\phi^* = f_{\theta^*}(I_m, I_f).
\]

The different choices of image similarity metrics and registration field regularizers \( (R(\phi)) \) were also studied in this work, and they are described in detail in the following subsections. The next subsection describes the design of ConvNet architecture.

### 1) ConvNet Architecture

From our experiments, we found that the choice of different network architectures does not have a large impact on the results. Our ConvNet was designed to be a U-Net-like “hourglass” architecture [32]. The network consists of one encoding path, which takes a single input formed by concatenating moving and fixed images into a \( 2 \times M \times M \) volume, where \( M \times M \) represents the shape of one image. Each convolutional layer has \( 3 \times 3 \) filter followed by a rectified linear unit (ReLU), and the downsampling was performed by the \( 2 \times 2 \) max pooling operations. In the decoding stage, the upsampling was done by "up-convolution" [32]. Each of the upsampled feature maps in the decoding stage, were concatenated with the corresponding feature maps from the encoding path. The outputting registration field, \( \phi \), was generated by the application of sixteen \( 3 \times 3 \) convolutions followed by two \( 1 \times 1 \) convolution to the 16 feature maps. The network architecture can be visualized in Figure 2.

#### 2) Image Similarity Metrics

Over the years, many research groups put considerable efforts into designing the image similarity metrics. We introduced some of the metrics that are broadly adopted in image registration in the last section. In this paper, we studied the effectiveness of four different loss functions, and later, we propose a new \( \mathcal{L}_{sim} \), which takes the advantage from the Pearson’s correlation coefficient and Structural similarity index (SSIM). In the following subsections, we denote the deformed moving image as \( I_d \) (i.e., \( I_d = I_m \circ \phi \)) for simplicity.

**a) Mean Squared Error (MSE)**

MSE is a simple measurement of how the intensity values line up between two images, it is applicable when \( I_f \) and \( I_m \) have similar contrast and intensity distributions. MSE is given by:

\[
\text{MSE}(I_d, I_f) = \frac{1}{\Omega} \sum_{i \in \Omega} \| I_f(i) - I_d(i) \|^2, \tag{5}
\]

where \( \Omega \) is the image domain. Then, the similarity loss function can be defined as \( \mathcal{L}_{sim}(I_m, I_f, \phi; \theta) = \text{MSE}(I_d, I_f) \).

**b) Pearson’s Correlation Coefficient (PCC)**

PCC measures the linear correlation between two images. Unlike MSE, PCC is less sensitive to the linear transformations of intensity values from one image to another. Its usage in medical image registration can be found in [33]. PCC is defined as the covariance between images divided by the product of their standard deviations:

\[
\text{PCC}(I_d, I_f) = \frac{\sum_{i \in \Omega} (I_f(i) - \bar{I}_f)(I_d(i) - \bar{I}_d)}{\sqrt{\sum_{i \in \Omega} (I_f(i) - \bar{I}_f)^2} \sqrt{\sum_{i \in \Omega} (I_d(i) - \bar{I}_d)^2}}, \tag{6}
\]

where \( \bar{I}_f \) and \( \bar{I}_d \) represents the mean intensities. PCC has a range from -1 to 1, where 0 implies that there is no linear correlation, and -1 corresponds to the maximum negative correlation between two images. Since a positive correlation...
is more desired, we can define the loss function to be:
\[
\mathcal{L}_{\text{sim}}(I_m, I_f; \phi, \theta) = 1 - \text{PCC}(I_m \circ \phi, I_f).
\]

c) Local Cross Correlation (CC)

Another popular image similarity metric is CC, for its robustness to intensity variations between images, it can be formulated as follows [16], [22], [34]:
\[
\text{CC}(I_d, I_f) = \frac{\left( \sum_{p \in \Omega} (I_f(p) - \bar{I}_f)(I_d(p) - \bar{I}_d) \right)^2}{\left( \sum_{p \in \Omega} (I_f(p) - \bar{I}_f)^2 \right) \left( \sum_{p \in \Omega} (I_d(p) - \bar{I}_d)^2 \right)},
\]

where \( I_d \) is the deformed image (i.e., \( I_d \)), \( p_i \) represents the pixel location within a window \( \Omega \), and \( \bar{I}_f \) and \( \bar{I}_d \) denote the local mean intensities within the window. Since \( \text{CC} \geq 0 \), we minimize the negative CC. Then, the loss function is \( \mathcal{L}_{\text{sim}}(I_m, I_f; \phi, \theta) = -\text{CC}(I_m \circ \phi, I_f) \).

d) Structural Similarity Index (SSIM)

SSIM was first introduced in [35] for robust image quality assessments based on the degradation of structural information. Within a given image window, SSIM is defined by:
\[
\text{SSIM}(I_d, I_f) = \frac{(2\mu_{I_d} \mu_{I_f} + C_1)(2\sigma_{I_d I_f} + C_2)}{\mu_{I_f}^2 + \mu_{I_d}^2 + \sigma_{I_f}^2 + \sigma_{I_d}^2 + C_2},
\]

where \( C_1 \) and \( C_2 \) are small constants for avoiding instability, \( \mu_{I_f} \) and \( \mu_{I_d} \), and \( \sigma_{I_f} \) and \( \sigma_{I_d} \) are local means and standard deviations of the image \( I_f \) and \( I_d \), respectively. SSIM has a range from -1 to 1, where 1 indicates a perfect structural similarity. Thus, \( \mathcal{L}_{\text{sim}}(I_m, I_f; \phi, \theta) = 1 - \text{SSIM}(I_m \circ \phi, I_f) \).

**C. Registration Procedure**

The algorithm for the proposed method is shown in Algorithm 1. For a given pair of moving and fix image, \( I_m \) and \( I_f \), an untrained ConvNet (\( f_0 \)) was initialized. First, the untrained \( f_0 \) produces an initial deformation field, \( \phi \). Second, we deform the moving image by \( \phi \) (i.e., \( I_m \circ \phi \)). Then, the loss between \( I_d \) and \( I_f \) is computed for the use of updating the parameters in the \( f_0 \). The above procedure is repeated until we hit the maximum number of iterations.

Since no information other than the given image pair is needed, the proposed method requires no previous training, thus it is fully unsupervised. The ConvNet is capable of learning an “optimal” deformation from a single pair of image. In the next section, we discuss the performance comparisons between this method and the state-of-the-art unsupervised methods.

**III. EXPERIMENTS**

We aim to create patient-like phantoms by registering the existing XCAT phantom with patient CT images. There were
Fig. 5: Comparisons between our method and the SyN algorithm from the ANTs package. The 1st column: moving image (XCAT phantom). The 2nd column: target image (patient CT scan). The 3rd column: deformed moving image by the proposed method with SSIM+CC loss. The 4th column: deformed moving image by SyN with MSE. The 5th column: SyN with CC. The 6th column: SyN with MI.

Fig. 6: Qualitative results from our method. The 1st row: target images. The 2nd row: deformed XCAT phantom. The 3rd row: deformed SPECT phantom.

Algorithm 1 ConvNet Registration

```
1: procedure CNNREG($I_m$, $I_f$) ▷ Input $I_m$ and $I_f$
2:     $f_\theta = \text{Initialize(ConvNet)}$
3: while $i < \text{iter}$ do ▷ For iter number of iterations
4:     $\phi = f_\theta([I_m, I_f], I_f)$ ▷ Predict deformation, $\phi$
5:     $I_d = I_m \circ \phi$ ▷ Deform moving image, $I_m$
6:     $\ell = \mathcal{L}(I_d, I_f; \theta)$ ▷ Compute loss
7:     $f_\theta = \text{BackPropagate}(f_\theta, \ell)$ ▷ Update ConvNet
8:     $i = i + 1$
9: return $I_d, \phi$
```

nine clinical low-dose whole-body CT patient scans used in this study; for those, only the torso part of the scans was extracted, which is 1153 2D-transaxial slices in total. The data was obtained from a publicly available dataset (NaF Prostate, [37]) in the Cancer imaging archive (TCIA, [38]). We first compare the performance generated by the ConvNets with different image similarity metrics. Then, we compare the proposed method with a state-of-the-art registration algorithm, the symmetric image normalization method (SyN) [16] from the ANTs package [39].

A. Loss Function Comparisons

Some examples of the registered XCAT phantom image by the five loss functions were shown in Figure. 3. (a) and (h) represent a same moving image, and (b) and (i) are the target images from the same CT slice, where the later was blurred...
by a low-pass Gaussian filter due to the presence of beam hardening artifacts. The third column to the last column are registration results by PCC, SSIM, PCC+SSIM, MSE, and CC, respectively. MSE and CC are two common loss functions in both traditional and learning-based image registration methods [14], [16], [18], [20]–[23], [34], but they failed to converge to good results (last two columns in Figure. 3). While PCC is robust to beam hardening artifacts, it produced an cartoonish contents around the spine (referring to (c) and (j)). On the other hand, SSIM completely models the noises and artifacts in the target image. The results produced by SSIM+PCC are much more balanced, combining with the Gaussian filter to additionally reduce noises, SSIM+PCC generated the best qualitative results among other loss functions.

| Method           | SSIM       | MSE       |
|------------------|------------|-----------|
| Affine only      | 0.828 ± 0.008 | 69.213 ± 2.748 |
| Ours             | 0.955 ± 0.007 | 37.340 ± 5.078 |
| SyN (MSE)        | 0.884 ± 0.011 | 51.999 ± 4.135 |
| SyN (MI)         | 0.881 ± 0.011 | 55.050 ± 3.996 |
| SyN (CC)         | 0.886 ± 0.011 | 52.838 ± 4.138 |

TABLE I: Comparison of SSIM and MSE between the proposed and the SyN method, where the best performances were highlighted.

B. Registration Performance Comparisons

In this experiment, we compared the proposed method with the SyN algorithm [16], Figure. 5 shows some qualitative comparisons between the proposed method and the SyN method. The first and the second columns indicate moving and fixed images. The third column shows the results generated by the proposed method. The fourth to the last columns represent the results obtained by SyN with MSE, CC, and MI, respectively. At a glance, our method gave a more accurate deformation than SyN, where the anatomy of bone structures and soft tissues were modeled precisely. Figure. 6 displays some qualitative results from the proposed method. The first row indicates the target images. The second and last rows show the deformed moving image and the deformed bone labels, respectively. Since the gold-standard is not available for the NaF Prostate dataset [37], the registration performances were evaluated quantitatively based on MSE and SSIM between $I_m \circ \phi$ and $I_f$. The results are shown in Table. I. The proposed method gave a mean SSIM of 0.955 and a mean MSE of 37.340, which outperformed the SyN method by a significant margin.

C. SPECT image simulations

Figure. 7 shows the results of mapping the XCAT phantom to a patient CT image. (a) and (b) exhibit the volume renderings of the phantom using 3DSlicer [40], (c) and (d) show a coronal slice of the deformed phantom and the SPECT simulation, respectively. SPECT projections were simulated by an analytic projection algorithm that realistically models attenuation, scatter, and the spatially-varying collimator-detector response [41], [42]. SPECT images were reconstructed using a subsets expectation-maximization algorithm (OS-EM) [43] based method [44], with 2 iterations and 10 subsets.

Fig. 7: Visualizations of deformed phantom and SPECT simulations. (a) Volume rendering of the deformed phantom. (b) Rendering of the skeleton. (c) A coronal slice of the phantom. (d) A coronal slice of the simulated SPECT image.

IV. Conclusion

This paper proposed to create patient-like phantoms with a ConvNet-based unsupervised and end-to-end registration technique that requires no prior training. Furthermore, we showed that the registration performance was significantly improved by combining SSIM and PCC as a data similarity loss function. The registration method was evaluated on the application of registering XCAT phantom with real patient CT scans and compared the registration performance in terms of SSIM and MSE to a state-of-the-art image registration method. Both quantitative and qualitative analysis indicate that our method provided the best results. Combined with Monte Carlo and CT simulation programs, the phantoms generated by our method are able to be transformed into more realistic human-like simulations.

ACKNOWLEDGMENT

This work was supported by a grant from the National Cancer Institute, U01-CA140204. The views expressed in written conference materials or publications and by speakers and moderators do not necessarily reflect the official policies of the NIH; nor does mention by trade names, commercial practices, or organizations imply endorsement by the U.S. Government.

We would like to show our gratitude to Dr. Daniel Tward and Shuwen Wei for sharing their pearls of wisdom with us during the course of this research.

REFERENCES

[1] C. P. Christoffersen, D. Hansen, P. Poulsen, and T. S. Sorensen, “Registration-based reconstruction of four-dimensional cone beam computed tomography,” IEEE Transactions on Medical Imaging, vol. 32, no. 11, pp. 2064–2077, 2013.
[39] B. B. Avants, N. J. Tustison, G. Song, and J. C. Gee, “ANTS: Open-source tools for normalization and neuroanatomy,” IEEE Transactions on Biomedical Engineering, vol. 10, pp. 1–11, 2009. [Online]. Available: ftp://ftp.heanet.ie/mirrors/sourceforge/a/ad/advants/Documentation/antstheory.pdf

[40] A. Fedorov, R. Beichel, J. Kalpathy-Cramer, J. Finet, J.-C. Fillion-Robin, S. Pujol, C. Bauer, D. Jennings, E. Feunessy, M. Sonka, J. Buatti, S. Aylward, J. V. Miller, S. Pieper, and R. Kikinis, “3d slicer as an image computing platform for the quantitative imaging network,” Magnetic Resonance Imaging, vol. 30, no. 9, pp. 1323 – 1341, 2012, quantitative Imaging in Cancer. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0730725X12001816

[41] E. C. Frey and B. M. W. Tsui, “A practical method for incorporating scatter in a projector-backprojector for accurate scatter compensation in spect,” IEEE Transactions on Nuclear Science, vol. 40, no. 4, pp. 1107–1116, Aug 1993.

[42] D. J. Kadrmas, E. C. Frey, and B. M. W. Tsui, “An svd investigation of modeling scatter in multiple energy windows for improved spect images,” IEEE Transactions on Nuclear Science, vol. 43, no. 4, pp. 2275–2284, Aug 1996.

[43] H. M. Hudson and R. S. Larkin, “Accelerated image reconstruction using ordered subsets of projection data,” IEEE Transactions on Medical Imaging, vol. 13, no. 4, pp. 601–609, Dec 1994.

[44] B. He, Y. Du, X. Song, W. P. Segars, and E. C. Frey, “A Monte Carlo and physical phantom evaluation of quantitative In-111 SPECT,” Physics in Medicine and Biology, vol. 50, no. 17, pp. 4169–4185, sep 2005.