Part - III

USING
LEARNING-BASED APPROACH
FOR
DEFORMABLE REGISTRATION

CHAPTER 6

EVALUATION OF A LEARNING BASED DEFORMABLE REGISTRATION METHOD ON ABDOMINAL CT IMAGES

Highlights of the Chapter

- *Deformable registration of abdominal images is done using deep learning techniques.*
- *Three experiments are done: 2D atlas-based, 2D pairwise, and 3D pairwise registration.*
- *Inter-subject registration results show competitive outcomes with reduced runtime.*

Abstract

Deformable registration of medical images generally relies on the iterative minimization of a cost function involving a large number of parameters. For complex deformations and large datasets, this process is computationally very demanding, leading to processing times that are incompatible with the clinical routine workflow. The recent breakthroughs in deep learning have led to revisit most of image processing problems, including deformable image registration. While building large annotated dataset can be a bottleneck for implementing such techniques, recent contributions demonstrated the potential of Convolutional Neural Networks (CNN) trained in an unsupervised way for achieving fast and efficient deformable registration of 2D and even 3D medical images. The present study proposes to evaluate one of the reference state-of-the-art methods VoxelMorph [5], in the context of inter-subject registration of abdominal CT images.

Experimental results obtained on two datasets, LiTS [6] and 3D-IRCADb-01 [7], highlight that VoxelMorph accuracy is comparable or even better than a reference non-learning based registration method ANTs (SyN) [12], with a drastically reduced computation time.

6.1. Introduction

Reliable image comparisons, based on fast and accurate deformable registration methods, are recognized as key steps in the diagnosis and follow-up of cancer as well as for radiation therapy planning or surgery [1]. In the particular case of abdominal images, the images to be registered often differ widely from each other. Differences arise, due to various reasons like organ deformation, patient motion, movements of gastrointestinal tract or breathing, yielding large local and global deformations that have to be compensated by the registration. As a consequence, the spatial correspondences which are to be established, are dense and highly non-linear. Moreover, the folding patterns in the deformation fields are complex and vary significantly between patients, making it difficult to learn simple parametric deformation models. Standard similarity measure-based approaches to deformable image registration amount to iteratively minimizing cost (objective) functions embedding thousand or millions of variables [2, 3, 4]. These registration algorithms are thus time-consuming, and typically require several minutes to hours for 3D images while running on the CPU. Difficulties may also arise due to the highly non-convex nature of these optimization problems.

In the last few years, deep learning-based frameworks have been introduced in deformable image registration to overcome these issues. After a computationally intensive offline learning on large training datasets, CNN-based registration methods are able to match a new pair of images within (fractions of) seconds. Among all the proposed approaches, the

VoxelMorph learning-based framework [5], made a major breakthrough since it does not require any ground truth deformation field for the learning phase, contrary to previous related approaches [15, 16, 17]. This is a crucial feature since it is extremely difficult to obtain accurate ground truth deformations from real data. An alternative strategy is to generate the learning datasets using synthetic warpings, but they may not necessarily be representative of the whole morphological variabilities that may be observed in real data. To overcome these limitations, VoxelMorph learns in an unsupervised way the complex mapping between every couple of 2D or 3D pairs of images and the corresponding deformation field by minimizing a standard intensity-based similarity metrics over the whole learning database. Any differentiable objective function can be considered. The training database simply consists of representative pairs of images. The mapping is parameterized using CNN. After the training step, evaluating the CNN output from an unseen pair of images enables us to obtain a deformation field within seconds. This learning-based approach allows, in a certain sense, to transfer the time consuming iterative minimizing procedure done by standard registration methods for each new pair of images to an offline computationally intensive learning phase, which needs to be carried out only once and for all.

In this paper, we present an evaluation of the reference CNN-based registration method VoxelMorph [5] in the context of inter-subject abdominal CT image registration by considering two datasets: LiTS [6] and 3D-IRCADb-01 [7]. Both image datasets are complemented with anatomical segmentations, which are used for evaluating the registration accuracy by calculating the overlap of each region of interest. VoxelMorph has so far only been evaluated on brain images. Abdominal images present a greater challenge in terms of registration than brain images,

due to greater anatomical variability and significant organ deformations. Our experimental results highlight that VoxelMorph performance is on a par with the classical non-learning-based state-of-the-art registration algorithm ANTs [4], but with a drastic reduction of the computation time to a fraction of seconds for 2D images and to a few seconds for 3D images which is more than two order of magnitude faster compared to ANTs.

The paper is organized as follows. In Section 6.2, the two databases of abdominal CTs used in this evaluation are described and the image registration method VoxelMorph is introduced. Section 6.3 is dedicated to the experimental results. Finally, conclusions are drawn in Section 6.4.

6.2. Material and Methods

6.2.1. Datasets

LiTS [6] and 3D-IRCADb-01[7] are the two datasets used for the experiments conducted in this work. The LiTS dataset contains 130 contrast-enhanced 3D abdominal CT scans and corresponding liver segmentation masks, which were created in collaboration with seven hospitals and research institutions and manually reviewed independently by three radiologists. The 3D-IRCADb-01 dataset is composed of 3D CT scans of 20 patients (10 men and 10 women) with the hepatic tumor in 75% cases. Masks of various segmented regions of interest (liver, spleen, skin, bone, kidney, etc) are also provided.

6.2.2. Background on Image Registration

Intensity-based image registration methods consist of warping a source or moving image I_m toward a target or fixed image I_f in order to put into correspondence pixels/voxels

corresponding to the same anatomical location. In general, deformable matching is preceded by a global affine image transform step that aims to compensate for patient positioning in the scanner, scale factors between different individuals and other global differences. Deformable registration estimates the residual global and local non-linear transformations, which require higher degrees of freedom. Standard similarity-based deformable registration algorithms consist of three components [9]:

- A parametric or non-parametric transformation model, which embeds between thousands and millions of variables.
- A similarity metric, which quantifies the resemblance between the moving and the fixed images.
- An optimization procedure, which minimizes a global objective function based on the similarity metric, by iteratively updating the transformation parameters [10].

The process of optimization can be summed up as follows:

$$\dot{\omega} = \arg max\, F_c \left(I_f,\, I_m,\, \omega \right) \tag{6.1}$$

$$\dot{\omega} = \arg max\, F_{sm} \left(I_f,\, I_m \circ \omega \right) \; + \; \lambda\, F_{reg}\, (\omega) \tag{6.2}$$

where ω is the registration field that maps coordinates of I_f to coordinates of I_m, F_c is the global cost (objective) function to be minimized. F_{reg} and F_{sm} are defined as the regularization and the similarity metrics respectively. Standard similarity metrics include mean square error, mean absolute error, cross-correlation, local cross-correlation, mutual information or normalized mutual information [11, 12]. λ is a regularization weighting parameter controlling the smoothness of the deformation field. The deformation ω is often defined by a displacement

vector field, but it can also be specified as a velocity vector field [13]. Standard registration approaches minimize the cost function F_c for each pair of images that have to be registered [14]. VoxelMorph relies on a different paradigm which consists in the estimation of global mapping, that associates to every couple of images (I_f, I_m) the corresponding deformation field ω. This mapping modeled as a CNN is learned once and for all by minimizing the cost F_c averaged on the whole training database.

6.2.3. CNN-based Registration: VoxelMorph

The block diagram of the image registration method VoxelMorph is shown in Fig. 6.1. The basic idea is to learn a registration function $D_\psi(I_f, I_m)$, that associates for every pair of input images of the training dataset the corresponding deformation field (ω), where (ψ) are the convolution kernels that parameterize the network architecture. Stochastic gradient descent optimization is used to get the optimal parameters for the final registration function by minimizing the cost function. As an output, a registration field (ω) is obtained such that $I_f(p)$ and $[I_m \circ \omega](p)$ hopefully correspond to the same anatomical locations (p is considered as pixel/voxel). During the testing phase, the fixed (I_f) and moving (I_m) image are evaluated as a new input of the network, which outputs the corresponding deformation field ω. Finally, the warped image $(I_m \circ \omega)$ is obtained using a spatial transformation network (STN) module. In this paper, I_f *and* I_m are considered as 2D or 3D images consisting of single-channel grayscale data. I_f *and* I_m are supposed to be globally aligned using affine registration as a preprocessing step (more details are provided in Section 6.3.1.1) Further details about the CNN architecture, the similarity metric and the STN are provided in the sequel.

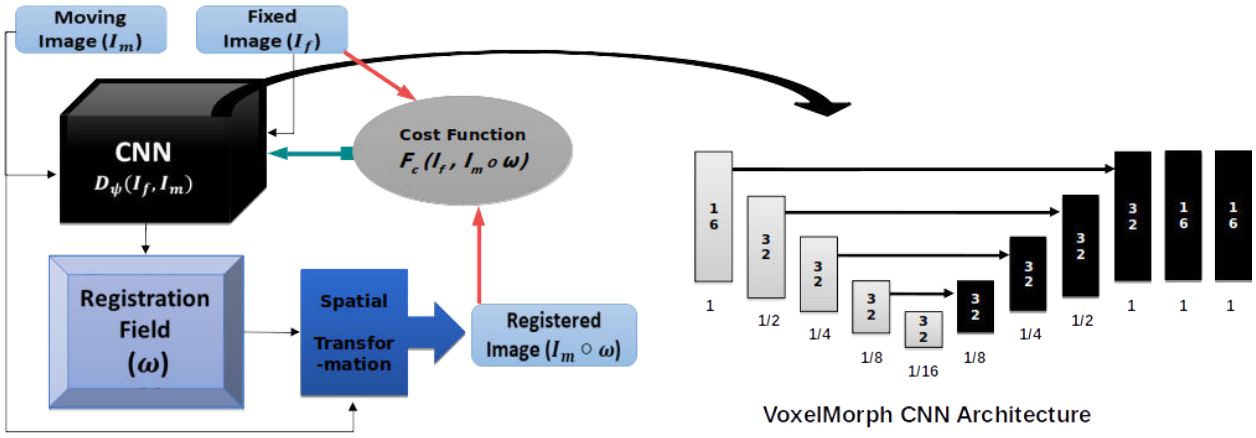


Fig. 6.1. Left side: Block diagram illustrating the VoxelMorph method
Right side: UNet Architecture considered for VoxelMorph CNN
Channel number and image spatial resolution are written inside and under the boxes, respectively

6.2.4 CNN Architecture

For the study and the experiments conducted in this paper, the VoxelMorph-2 network architecture is used [5]. The CNN considered for the parameterization of the registration function D_ψ is similar to the standard UNet network [18], which is divided into an encoder and decoder portions with skip connections between them (see Fig. 6.1).

The fixed and moving images are concatenated into a 2-channel image and given as input to the network. In both the encoder and decoder sections, convolutions are applied at each layer with kernels of size 3^n (n being the image dimension). These layers capture the image features and finally encode the deformation field (ω). A stride of 2 is used in the encoder part to decrease the image dimension into half of its previous layer's input to get the image features from its fine to coarse representation. A LeakyReLU layer with a multiplier of 0.2 is applied after each convolutional layer. In the decoding part, images are upsampled and concatenated to the images

of the same spatial resolution from the encoder stage through skip connections before convolutions. These skip connections are an important part of the architecture, as the features learned all over the encoder stage layers directly propagate to the decoder stage layers. The features convolve on successive coarse to fine spatial resolution and help generate precise deformation field (ω) for anatomical alignment.

6.2.5. Cost Function

Any differentiable cost function can be considered within the VoxelMorph framework. As previously described, the cost function is composed of two terms, the similarity metric F_{sm}, and the regularization function F_{reg}:

$$F_c\left(I_f,\ I_m,\ \omega\right) = F_{sm}\left(I_f,\ I_m \circ \omega\right) + \lambda F_{reg}(\omega) \tag{6.3}$$

In this study, two similarity metrics are considered: the mean square error (MSE) and the local cross-correlation (CC). MSE is basically defined by the squared pixelwise/voxelwise differences of the two images:

$$MSE\ \left(I_f,\ I_m \circ \omega\right) = \frac{1}{|\Omega|} \sum_{p \in \Omega} \left[I_f(p) - [I_m \circ \omega](p)\right]^2 \tag{6.4}$$

where the images are defined over the spatial domain Ω. This similarity metric requires that both images have similar intensity distributions. To cope with images that have different contrast, local cross-correlation (CC) can advantageously be used:

$$CC\ \left(I_f,\ I_m \circ \omega\right) = \frac{\left[\sum_{p_i}\left(\left(I_f(p_i) - \hat{I}_f(p)\right)\left([I_m \circ \omega](p_i) - \left[\hat{I}_m \circ \omega\right](p)\right)\right)\right]^2}{\left(\sum_{p_i}\left(I_f(p_i) - \hat{I}_f(p)\right)^2\right)\left(\sum_{p_i}\left([I_m \circ \omega](p_i) - \left[\hat{I}_m \circ \omega\right](p)\right)^2\right)} \tag{6.5}$$

where $\hat{I}_f(p)$ and $\left[\hat{I}_m \circ \omega\right](p)$ designate the local mean intensities images computed on a local neighborhood ($9 \times 9 \times 9$ window is considered in our experiments). Minimizing the cost function solely on the basis of the similarity measure can end up to a physically unrealistic non-smooth deformation field (ω). Hence, an additional regularization term is considered that penalizes non-smooth deformation fields (ω) based on the norm of the spatial derivative of the deformation field:

$$F_{reg}(\omega) = \sum_{p \in \Omega} \| \nabla \omega(p) \|^2 \tag{6.6}$$

This regularization term is introduced in the cost function to achieve anatomically consistent results by preventing sharp discontinuities in the estimated deformation field. Its influence is modulated through the weighting parameter λ. Large values of λ (in equ. 3) will favor smooth, locally constant deformation fields, whereas values close to zero increases the risk of observing foldings and sharp discontinuities in the estimated deformation field. The influence of λ is evaluated and discussed in Section 3.5.

6.2.6. Spatial Transformation Function

To compute the warped image ($I_m \circ \omega$), a spatial transformation module implementing linear interpolation is considered. Using this differentiable operator allows the backpropagation of errors during the optimization process. More precisely, for every pixel/voxel p, I_m is evaluated at a subpixel/subvoxel location $\omega(p)$ using trilinear interpolation by considering the four (resp. eight) neighboring pixels (resp. voxels) $N(\omega(p))$.:

$$I_m \circ \omega\,(p) = \sum_{q \in N(\omega(p))} I_m(q) \prod_{d \in \{x,y,z\}} \left(1 - \left|\omega_d\,(p) - q_d\right|\right) \tag{6.7}$$

where d iterates over the dimension of Ω, i.e., $d \in \{x,y\}$ (resp. $d \in \{x,y,z\}$) for 2D (resp. 3D) images.

6.3. Experiments and Results

The experiments have been performed on two databases of abdominal CT images. VoxelMorph is benchmarked with the competitive deformable registration method Symmetric Normalization (SyN) [12] available through the software package of ANTs (http://stnava.github.io/ANTs/). The registration methods are evaluated for both 2D and 3D image registration. In this section, we first describe the experimental setup including the preparation of the dataset and the metrics used for the evaluation. Then, results are presented focusing on three scenarios, namely 2D registration on a reference image (considered as an atlas), 2D pairwise registration (without a reference image) and 3D pairwise registration (without a reference image). The influence of the weighting regularization parameter is finally investigated.

6.3.1. Experimental Setup

Experiments are conducted with the implementation of VoxelMorph (https://github.com/voxelmorph/voxelmorph) using Keras and Tensorflow as a backend. The ADAM optimizer is used with a learning rate of 10^{-3}. A mini batch stochastic gradient descent technique is employed for every experiment with a batch size of 4. Training steps are executed

on an NVIDIA GeForce GTX 1080 GPU (8 GB), while all the testing steps are computed using an Intel Core-i7 3770 CPU (3.40 Hz).

6.3.1.1. Dataset Preparation

Abdominal 3D CT-scan volumes from the LiTS [6] and 3D-IRCADb-01 [7] dataset are used for the assessment of VoxelMorph. As the LiTS dataset has more 3D volumes than 3D-IRCADb-01 (130 against 20), only the volumes from LiTS are taken into account for training the CNN, while the volumes from both datasets are considered for the evaluation.

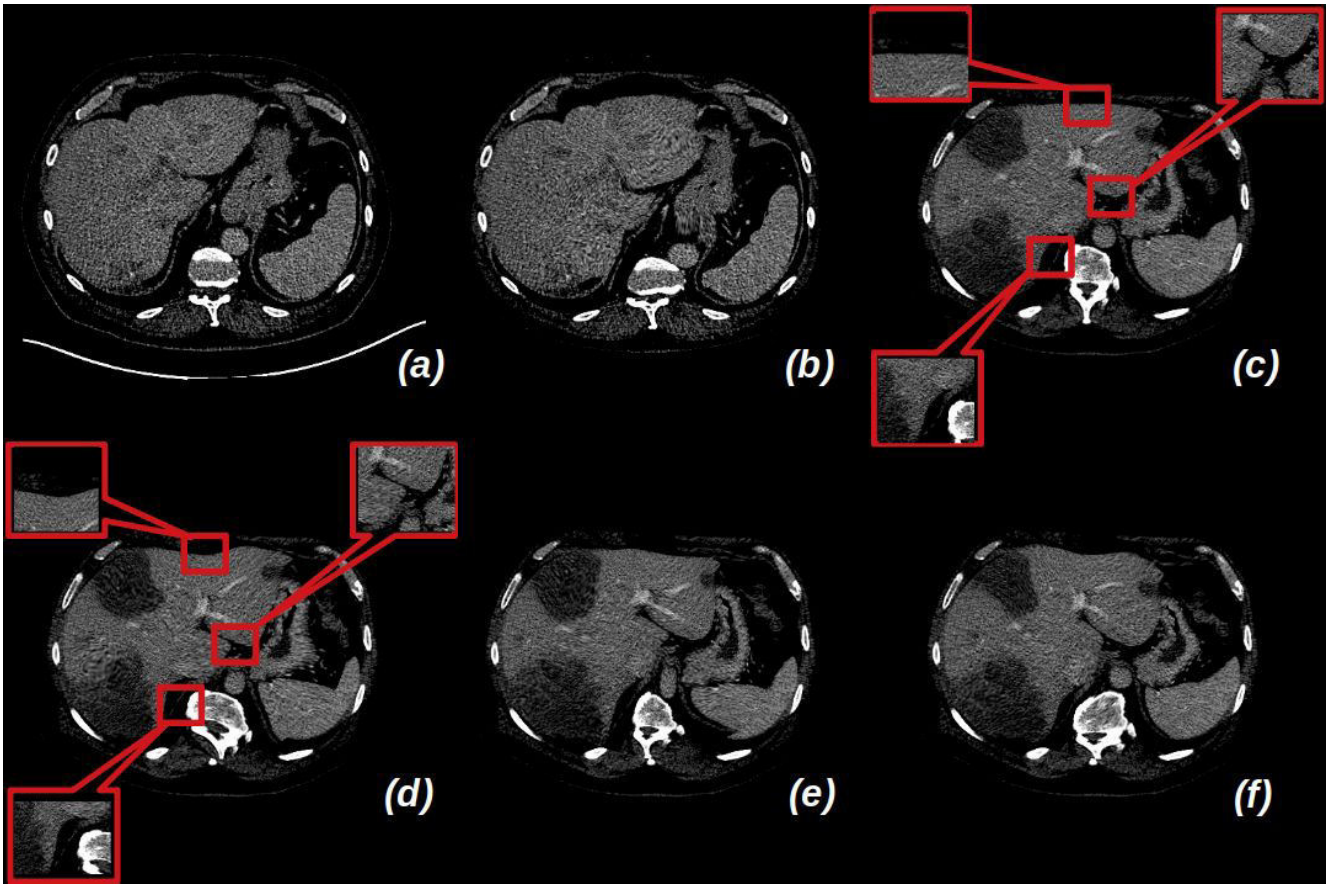


Fig. 6.2. (a) original 2D abdominal image; (b) 2D abdominal image after bed removal; (c) original 2D abdominal image and (d), (e), (f) examples of warped images generated with Gryds package for data augmentation.

As the main focus of this study is the deformable registration of abdominal organs, a bounding box including the liver and some portion above and below is cropped out from the 3D volumes of both datasets. Several preprocessing are then applied to the images before the training phase. Two different workflows are considered for 2D images and 3D volumes, respectively.

- **2D dataset**

This dataset was created by extracting 2D slices from the cropped 3D volumes. Firstly, the 3D volume intensities are clipped between 0 and 255. Artefacts in the volume image (structures other than the abdominal portion, e.g. patient bed) are removed using morphological operators (see Fig. 6.2.(a)-(b)). Then a random 2D axial slice of size (512 × 512) is chosen as the reference 2D image or atlas. Every slice of the cropped 3D volumes are aligned with the 2D atlas image by affine transformation (with the help of the ANTs-registration suite), and the best matching 2D slice (evaluated through mean absolute error) from each 3D volume is considered for further processing. Then, the histogram matching technique implemented in ANTs is applied to normalize the 2D images. As a result, 81 affinely aligned 2D images and a reference image are obtained. To increase the number of images for training, data augmentation is performed using the Gryds [19] package (https://github.com/tueimage/gryds). Gryds implements deformable transformations modeled as B-splines. It has been modified as per the requirement of this study. The control points on the ribs regions of the abdominal images have been constrained to remain stationary, while the control points inside the abdominal portion are allowed to move (shown in Fig. 6.2.(d), (e), (f)). This constraint is introduced to obtain anatomically plausible deformed images. Finally, a dataset has been produced, with 810 abdominal 2D images of size (512 × 512) with a 1 mm isotropic pixel dimension. The dataset has been split into 730 and 80 images for

training and testing respectively. The same dataset is used for 2D atlas-based registration and 2D pairwise registration.

- **3D dataset**

For the preprocessing of the 3D volumes, again the first step is to clip the volume intensities between 0 and 255 and to discard structures other than the abdominal portion. The height and the width of the cropped 3D volumes are all (512 × 512), but the depth of each volume is different. All cropped 3D volumes are thus resampled to (512 × 512 × 256) with 1mm isotropic voxels. Next, one random volume from the cropped resampled 3D volumes of the LiTS dataset is considered as a reference volume, and intensity normalization is carried out for all other volumes using the histogram matching technique implemented in ANTs. Finally, an affine transformation is used to align the 3D volumes with the reference volume (using ANTs-Registration). As a result, 111 abdominal 3D volumes of size (512 × 512 × 256) with 1mm isotropic voxel dimension are obtained. This dataset has been split into 100 and 11 images for training and testing respectively, for 3D pairwise registration. Five out of 20 images from the 3D-IRCADb-01 dataset, are processed in the same way and considered as testing data. This subset corresponds to the images for which the segmentations of bone, kidneys, liver, portal vein and skin were all provided.

6.3.1.2. Evaluation Metric

In image registration, the pixels/voxels with similar appearances are aligned in order to put into correspondence anatomical structures. Consequently, quantifying the overlap and measuring the point to point distances between segmentation maps of several anatomical

structures, associated with the moving image and the fixed image, is one of the possible strategies generally used for evaluating the accuracy of the registration.

- **Dice Score**

The Dice score has been used here to quantify the overlap between two binary segmentation maps for both 2D and 3D images. If the anatomical structures of I_f and $I_m \circ \omega$ overlap well, the Dice score is close to 1, while no overlap leads to a Dice score of 0. For each anatomical structure, the segmentation maps of the moving image are warped using the deformation fields obtained with the different registration methods. The Dice score is defined as:

$$Dice\left(R^s_{I_m \circ \omega},\ R^s_{I_f}\right) = 2 * \left[\frac{R^s_{I_m \circ \omega} \cap R^s_{I_f}}{\left|R^s_{I_m \circ \omega}\right| + \left|R^s_{I_f}\right|}\right] \tag{6.8}$$

where $R^s_{I_m \circ \omega}$ *and* $R^s_{I_f}$ are the set of pixels/voxels representing a particular anatomical structure s in the registered and fixed images respectively.

- **Hausdorff Distance**

The Hausdorff distance quantifies how far two subsets of points (here, two binary segmentation maps) are according to a given metric (here, the Euclidean norm is considered). More concretely, it is defined as the greatest of all the distances from a point in the segmentation map of I_f to the closest point in the warped segmentation map associated to $I_m \circ \omega$ and conversely. This distance reflects the worst degree of mismatch between two segmentation maps. Defining the two segmented anatomical structures $R^s_{I_f} = \{a_1,\ a_2, ..., a_p\}$ and $R^s_{I_m \circ \omega} = \{b_1,\ b_2, ..., b_q\}$ as two finite point sets, the Hausdorff distance is defined as:

$$HD\left(R^s_{I_f},\ R^s_{I_m\circ\omega}\right) = max\left(hd\left(R^s_{I_f},\ R^s_{I_m\circ\omega}\right), hd\left(R^s_{I_m\circ\omega}, R^s_{I_f}\right)\right) \quad (6.9)$$

where,

$$hd\left(R^s_{I_f},\ R^s_{I_m\circ\omega}\right) = max_{a\in R^s_{I_f}}\ min_{b\in R^s_{I_m\circ\omega}}\ ||a-b|| \quad (6.10)$$

The Dice Score and the Hausdorff Distance are calculated in the experiments on 2D atlas-based registration, 2D pairwise registration, and 3D pairwise registration for the liver region only, since this is the only segmented structure provided in the LiTS dataset. Since tissue label overlap scores may not be a reliable registration accuracy metric, especially for large structures such as the liver [8], we additionally considered smaller anatomical structures such as the kidney, and even more localized structures such as the portal vein provided in 3D-IRCADb-01 dataset for evaluating the model on 3D pairwise registration.

6.3.1.3 Comparison with a State-of-the-Art Method

In this study, the accuracy of VoxelMorph is compared with the Symmetric Normalization (SyN) method [12], which is a competitive method that has been proven to be efficient on several image modality and organs, such as brain MRI [3,4] or thoracic CT scan [20]. SyN is an implementation of the large deformation diffeomorphic metric mapping (LDDMM) algorithm, available in the ANTs software package. We adopted a step size of 0.25 and used cross-correlation (CC) as the similarity metric. The CC metric available in ANTs is a fast implementation of the standard normalized cross-correlation metric computed in a local neighborhood. The CC metric is known to result in registrations that are robust to differences in intensity contrast. We adopted a local neighborhood of size 3. More information about the different similarity metrics available in ANTs may be found in [4]. The registration method was

implemented for 4 multi-resolution steps with a maximum number of iterations of 500 per step to make sure that convergence is reached.

6.3.2. 2D Atlas-based Registration

In this experiment, the CNN model is trained with 730 2D affine registered abdominal images and one atlas image using mean square error (MSE) as the similarity metric. 80 images are kept as testing images for evaluating the performance of the method. The images are downsampled to (256 × 256) from their original size (512 × 512) to reduce the amount of computation and the training time as well. At a time, the atlas image and one training image are concatenated together and given as input to the model in batches. The total number of epochs are 600 and 700 steps are executed per epoch for the training. The value used for the regularization parameter is 0.25. The training of the model is done on an NVIDIA GeForce GTX 1080 (GPU) and the total training time for the 2D atlas-based registration experiment was 4.84 hours.

The average Dice scores and the Hausdorff distance metrics computed for the liver mask are reported for Affine, VoxelMorph, and ANTs (SyN with CC metric) in Table 6.1. As expected, both VoxelMorph and ANTs methods show significantly better performance than the affine alignment (higher dice scores and lower value of the Hausdorff distance). Fig. 6.3 shows examples of 2D atlas-based registrations. VoxelMorph performance is comparable to ANTs (SyN) in this case. A significant improvement is achieved in terms of computation time (Table 6.1), while considering an Intel Core-i7 3770 CPU for the online calculations. On average, VoxelMorph is more than 120 times faster than ANTs (SyN).

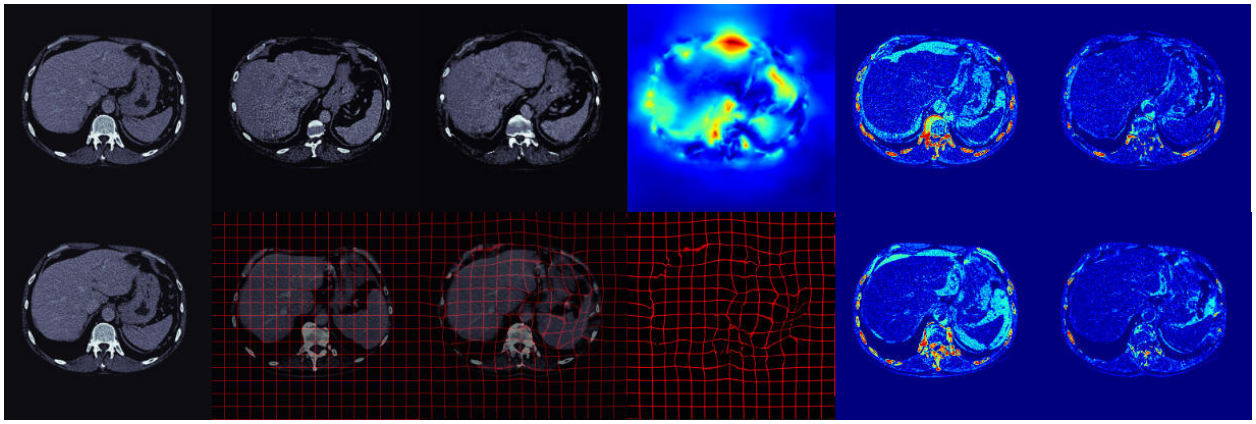


Fig. 6.3. Example of 2D atlas-based registration.

For images in first row: Dice Score (affine registration) 0.852; Dice Score (VoxelMorph) 0.94; Hausdorff Distance (affine registration) 5.291; Hausdorff Distance (VoxelMorph) 4.690

For images in second row: Dice Score (affine registration) 0.8812; Dice Score (VoxelMorph) 0.9355; Hausdorff Distance (affine registration) 4.472; Hausdorff Distance (VoxelMorph) 4.243

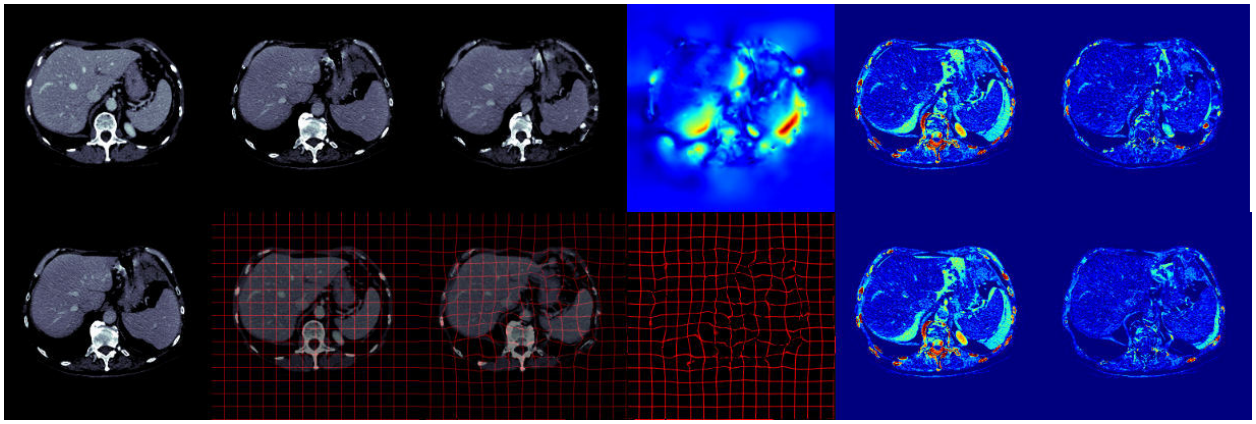


Fig. 6.4. Example of 2D Pairwise Registration

For images in first row: Dice Score (affine registration) 0.7962; Dice Score (VoxelMorph) 0.9019; Hausdorff Distance (affine registration) 3.873; Hausdorff Distance (VoxelMorph) 3.742;

For images in second row: Dice Score (affine registration) 0.8967; Dice Score (VoxelMorph) 0.9406; Hausdorff Distance (affine registration) 4.795; Hausdorff Distance (VoxelMorph) 4.013

6.3.3. 2D Pairwise Registration

Similarly to the previous experiment, 730 2D affine registered abdominal images are used as the training dataset (in this case without the atlas image), and eight images are kept for testing the accuracy (after discarding the rest 72 images augmented from those eight images and only considering the original ones). Again the similarity metric used is MSE and the images are downsampled to (256 × 256). In this experiment, pairs of images are randomly selected at a time and concatenated together to be used as input in batches for the model. Here 3600 steps per epoch are utilized while the number of epochs is 750. The regularization parameter is set to 0.25. It took 24.33 hours to train the model with an NVIDIA GeForce GTX 1080 GPU.

The average Dice scores and the Hausdorff distances for the liver mask obtained using Affine, VoxelMorph, and ANTs (SyN) are listed in Table 6.2. The metrics are computed from all combinations of pairwise registration using the eight testing images, corresponding to 56 test registrations. As in the previous experiment, the two deformable registration methods outperform the affine registration by a significant margin. Examples are shown in Fig. 6.4. It is interesting to notice that the performances of ANTs and VoxelMorph are still comparable in that experiment, while one could have expected a significant alteration of the performance of VoxelMorph in that scenario. Indeed, the pairwise registration scenario is far more challenging than the atlas-based registration scenario for VoxelMorph since the learned function $D_\psi(I_f, I_m)$ is now defined on a very larger domain (I_f is fixed in the atlas-based registration scenario while I_f is a variable in the pairwise registration scenario). This illustrates the ability of VoxelMorph to model very complex registration function $D_\psi(I_f, I_m)$. In terms of computational burden, VoxelMorph still produces output on average 120 times faster than ANTs (SyN).

Table 6.1. Quantitative Results (2D atlas-based abdominal image registration)

	Dice Score Mean (± SD)	Hausdorff Distance Mean (± SD)	Single Test Time (sec.)	Total Test Time (sec.)
Affine	0.8523 (0.0279)	4.6120 (0.5547)	---	---
VoxelMorph	0.9341 (0.0144)	4.2984 (0.4756)	**0.3551**	**9.9032**
ANTs (SyN)	**0.9431 (0.0175)**	**4.1973 (0.6419)**	15.199	1215.9

Table 6.2. Quantitative Results (2D pairwise abdominal image registration)

	Dice Score Mean (± SD)	Hausdorff Distance Mean (± SD)	Single Test Time (sec.)	Total Test Time (sec.)
Affine	0.8538 (0.0293)	4.8054 (0.7614)	---	---
VoxelMorph	0.9257 (0.0206)	**4.4555 (0.6917)**	**0.3508**	**6.7963**
ANTs (SyN)	**0.9318 (0.0266)**	4.4978 (0.7644)	15.2511	841.1878

Table 6.3. Quantitative Results (3D pairwise abdominal image registration; model using MSE)

	Dice Score Mean (± SD)	Hausdorff Distance Mean (± SD)	Single Test Time (sec.)	Total Test Time (sec.)
Affine	0.74929 (0.478)	23.488 (2.4824)	---	---
VoxelMorph	0.8352 (0.047)	20.3401 (3.039)	**6.6356**	**690.5838**
ANTs (SyN)	**0.8426 (0.064)**	**19.6499 (3.887)**	306.373	33701.08

Table 6.4. Quantitative Results (3D pairwise abdominal image registration; model using CC)

	Dice Score Mean (± SD)	Hausdorff Distance Mean (± SD)	Single Test Time (sec.)	Total Test Time (sec.)
Affine	0.74929 (0.478)	23.488 (2.4824)	---	---
VoxelMorph	0.8313 (0.0475)	20.4438 (2.889)	**7.1905**	**750.8254**
ANTs (SyN)	**0.8426 (0.064)**	**19.6499 (3.887)**	306.373	33701.08

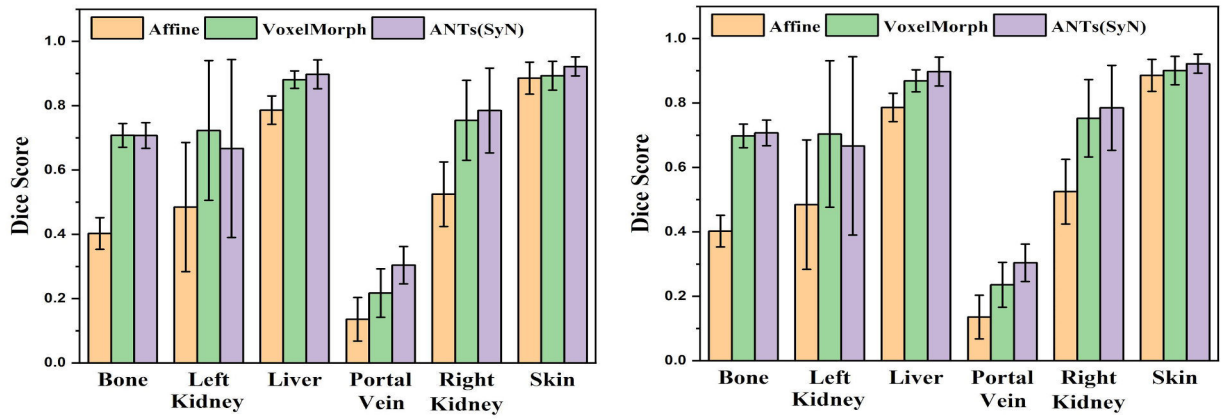


Fig. 6.5. Dice Scores: Mean (± SD); 3D-IRCADb-01; model trained using (left) MSE, (right) CC

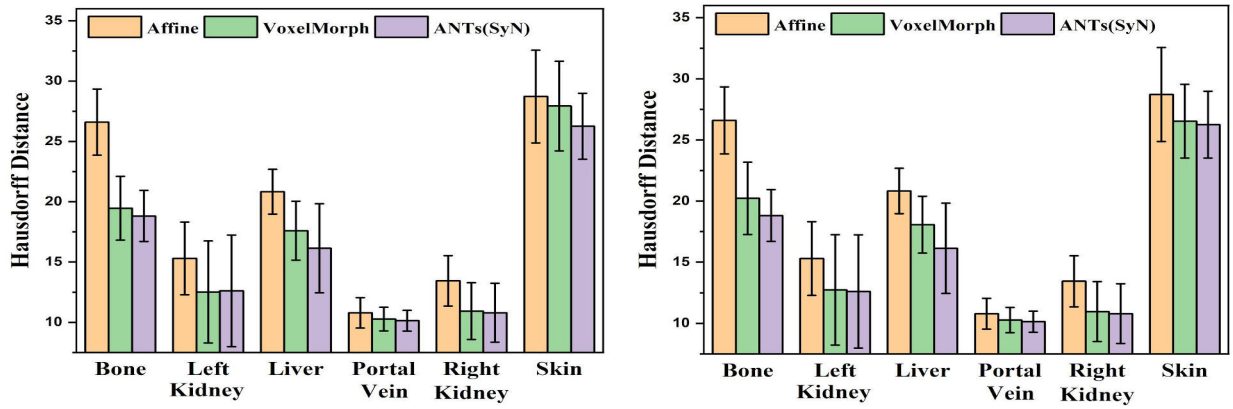


Fig. 6.6. Hausdorff Distance: Mean (± SD); 3D-IRCADb-01; trained using (left) MSE,(right) CC

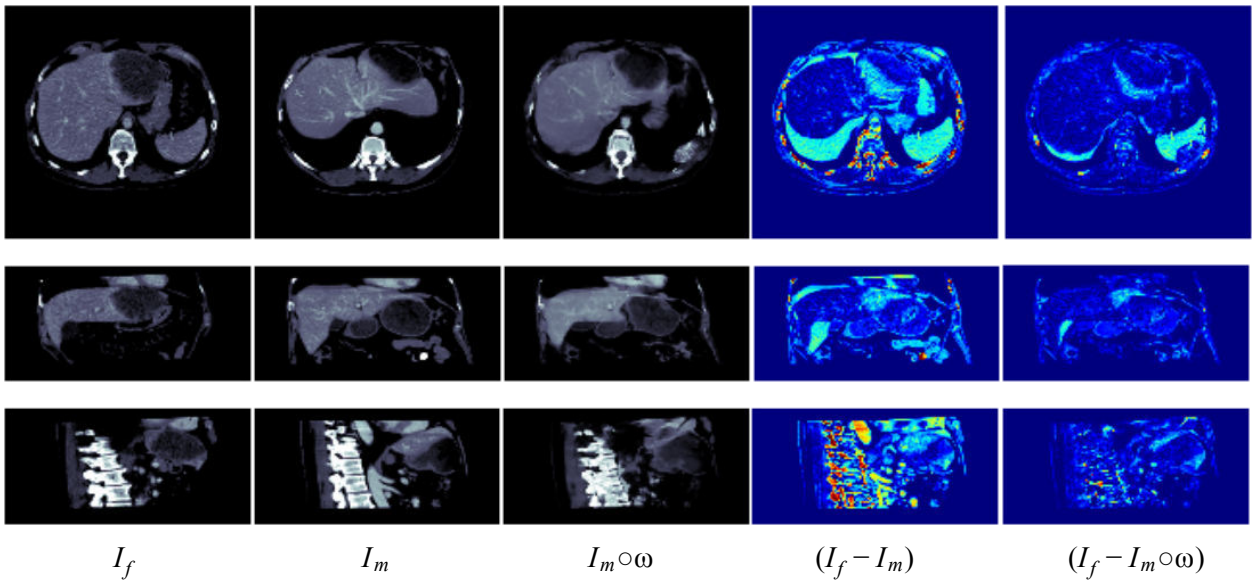


Fig. 6.7. Example of 3D Pairwise Registration (using MSE);
Dice Score (affine registration) 0.7093; Dice Score (VoxelMorph) 0.7955
Hausdorff Distance (affine registration) 27.9464; Hausdorff Distance (VoxelMorph) 21.7715

6.3.4. 3D Pairwise Registration

For this experiment, the training dataset consists of 100 3D affine registered abdominal images from the LiTS dataset. The original cropped images (512 × 512 × 256) are downsampled to (128 × 128 × 64) in order to contain the computational cost so that the training of the network can be performed faster while using only a single GPU, and test results can be computed using CPU only (no need to use GPU, hence memory allocation time is saved). Here also, two random images from the training dataset are picked up, concatenated and given as input in batches to train the model. The model is trained with two different similarity metrics, namely MSE and CC, using two appropriate values for the regularization parameter, 0.25 and 1.5 respectively. Using the NVIDIA GeForce GTX 1080 GPU, it takes around 11.023 hours to train the model with MSE (number of epochs = 75, steps per epoch = 600), while using CC, the time is about 19.83 hours (number of epochs = 90, steps per epoch = 600). 11 images from LiTS dataset (Fig. 6.7 and 6.8) and 5 images from 3D-IRCADb-01 dataset (Fig. 6.9) are used to evaluate the accuracy.

Table 6.3 and 6.4 summarize the Dice scores and the Hausdorff distances calculated for Affine, ANTs (SyN) and VoxelMorph based on MSE and CC similarity metrics. The values listed are computed from all possible combinations of pairwise registering the 11 test images of the LiTS dataset (110 test registrations in total). Similarly, Fig. 6.5 and Fig. 6.6 shows the mean Dice scores and the Hausdorff distance values respectively, obtained on the 3D-IRCADb-01 dataset for six anatomical regions of the abdomen, namely, bone, left kidney, liver, portal vein, right kidney, and skin. Only five out of 20 images were associated with the segmentation of all these six regions and were considered as test images, leading to 20 registration tests. The results indicate a definite improvement of the Dice scores and a precise drop of the Hausdorff distances

after using VoxelMorph or ANTs (SyN). It is observed that the output of VoxelMorph is slightly worse than ANTs (SyN). In this 3D registration case study, VoxelMorph is almost 50 times faster than ANTs (SyN). 3D registration with VoxelMorph takes only a few seconds using the CPU. So the slight loss in performance is therefore deliberately compensated for by the CPU time savings.

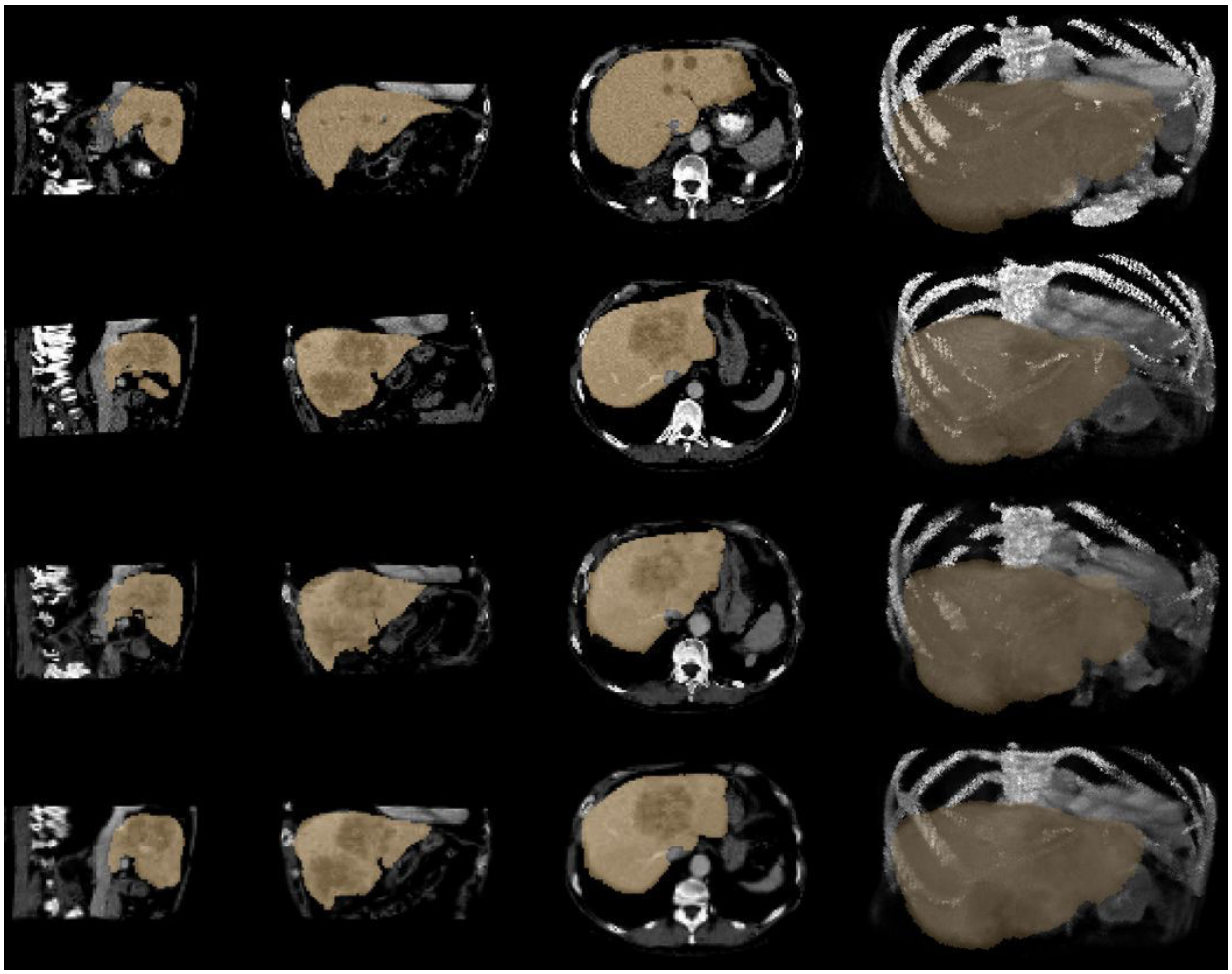


Fig. 6.8. LiTS volume registration with segmentation overlapping; Triplanar view and 3D view
1st row: I_f ; 2nd row: I_m ; 3rd row: $I_m \circ \omega$ (using VoxelMorph);
4th row: $I_m \circ \omega$ (using ANTS(SyN))
for Affine: Dice Score: 0.7941 ; Hausdorff Distance: 20.832
for VoxelMorph: Dice Score: 0.881 ; Hausdorff Distance: 19.261
for ANTS(Syn): Dice Score: 0.8981 ; Hausdorff Distance: 18.973

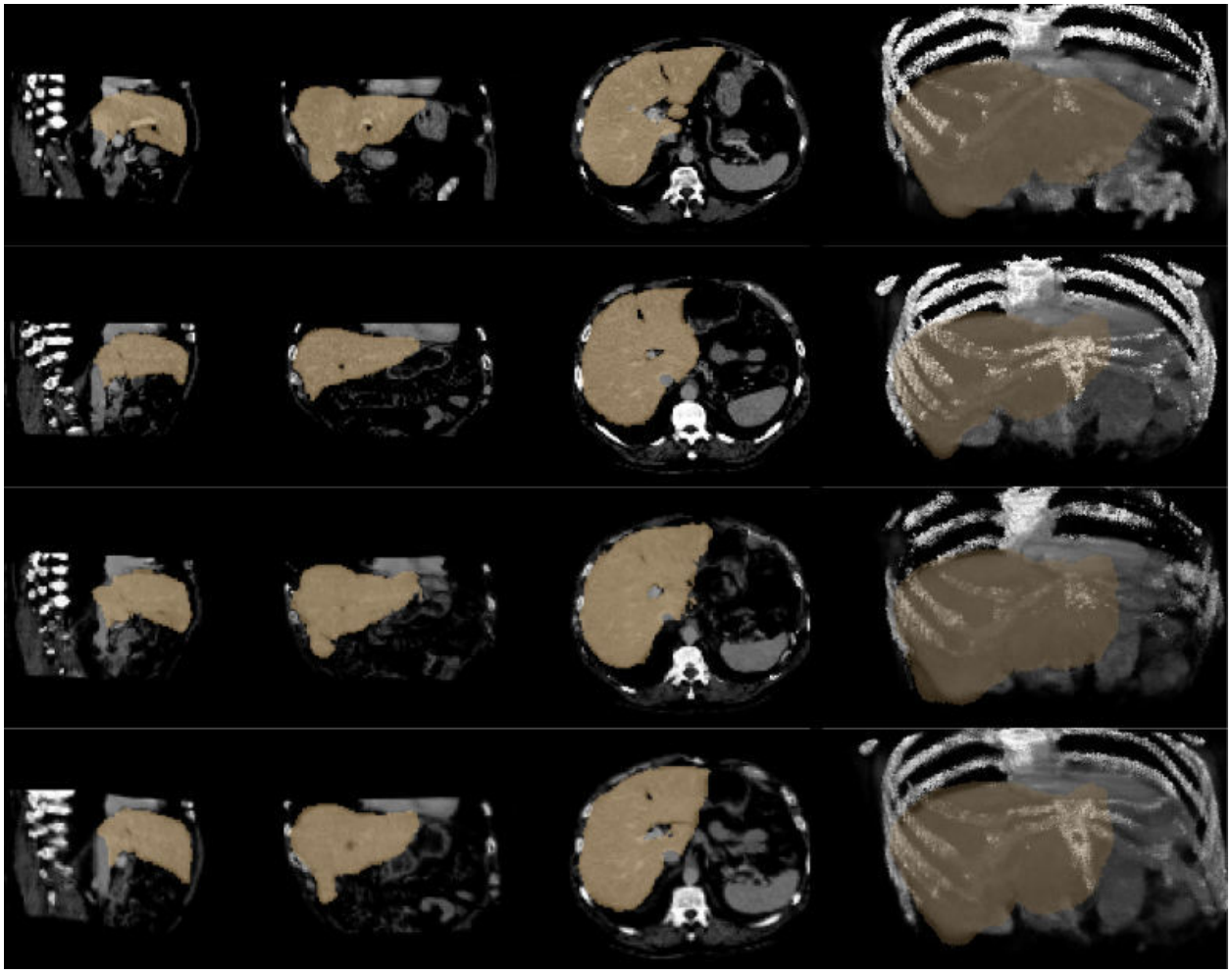


Fig. 6.9. 3D-IRCADb-01 vol. registration with segmentation overlapping; Triplanar & 3D view;
1st row: I_f ; 2nd row: I_m ; 3rd row: $I_m \circ \omega$ (using VoxelMorph);
4th row: $I_m \circ \omega$ (using ANTS(SyN))
for Affine: Dice Score: 0.8084 ; Hausdorff Distance: 21.071
for VoxelMorph: Dice Score: 0.9079 ; Hausdorff Distance: 15.394
for ANTS(Syn): Dice Score: 0.9126 ; Hausdorff Distance: 14.817

6.3.5. Setting of the Regularization Parameter

The quality of image registration is known to depend on the smoothness term in the cost function, the effect of which is controlled by the regularization parameter λ. The test dataset from LiTS data is evaluated on 2D atlas-based registration, with varying regularization

parameters ranging from 0 to 1 and the average Dice Scores are calculated (Fig. 6.10). As the graph shows, the best score is achieved when the regularization parameter is set to 0.25. So, for the other experiments with 2D abdominal images, when MSE is used as the similarity metric, λ is set to 0.25. Surprisingly, the graph shows a significant improvement of the Dice score from the affine registration, while the λ is set to zero, which means no smoothness is applied. This clearly indicates that an embedded smoothing capability is generated in the model function when it is trained to register images. In a similar way, the regularization parameter λ is set to 1.5 when local CC is used as the similarity metric.

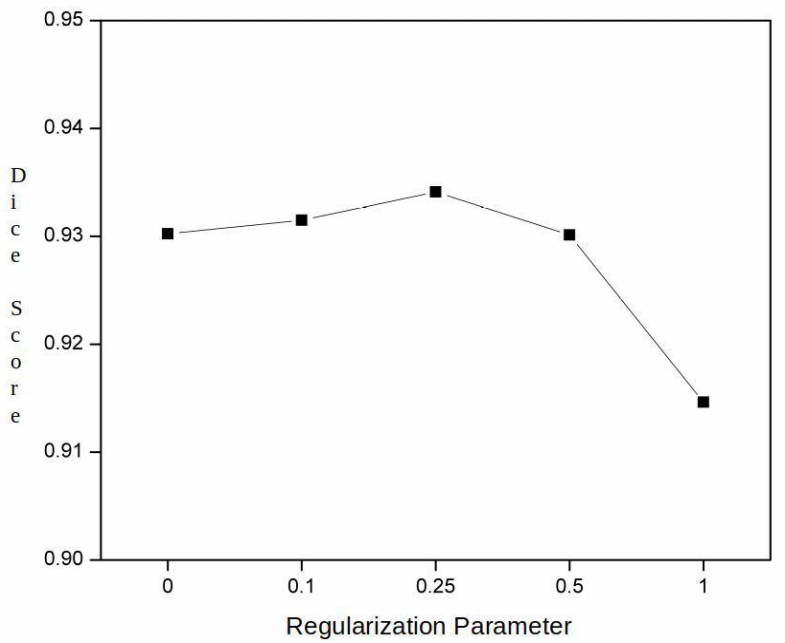
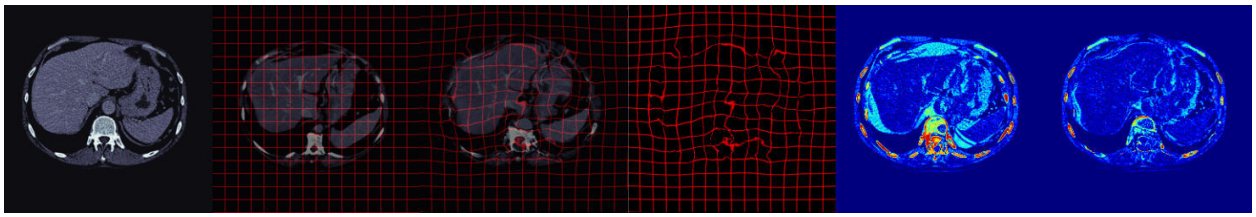


Fig. 6.10. Influence of the regularisation weighting parameter, for 2D Atlas-based Registration

6.4. Discussion

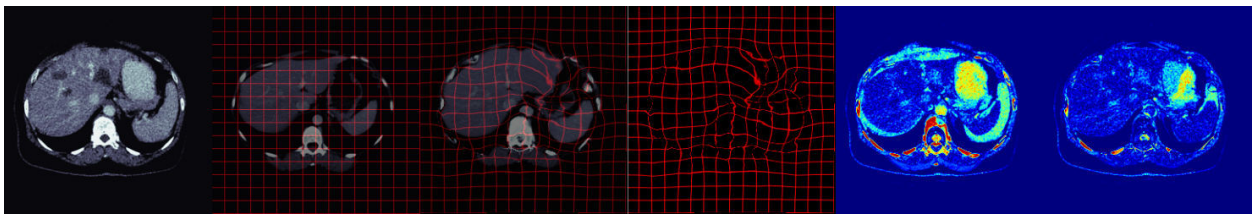
In this study, we have proposed an evaluation of VoxelMorph, a state-of-the-art unsupervised learning-based framework for deformable image registration on 2D and 3D abdominal CT-scan images. The performances of VoxelMorph were compared with a

competitive non-learning-based deformable registration method "Symmetric Normalization" (SyN), implemented in ANTs, on two representative databases: LiTS and 3D-IRCADb-01. Three different experiments were carried out on 2D or 3D data, on atlas-based or pairwise registration, using two different similarity metrics, namely MSE and CC. Accuracy of the registration was measured by the Dice score and Hausdorff distance metrics, which respectively quantify the volume overlap and the farthest point to point distance for selected anatomical regions.



I_f I_m $I_m \circ \omega$ ω $(I_f - I_m)$ $(I_f - I_m \circ \omega)$

Fig. 6.11. Example of 2D Atlas-based Registration (poor case for VoxelMorph)
Dice Score (affine registration) 0.78699; Dice Score (VoxelMorph) 0.8949;
Hausdorff Distance (affine registration) 5.099; Hausdorff Distance (VoxelMorph) 4.472



I_f I_m $I_m \circ \omega$ ω $(I_f - I_m)$ $(I_f - I_m \circ \omega)$

Fig. 6.12. Example of 2D Pairwise Registration (poor case for VoxelMorph)
Dice Score (affine registration) 0.83547; Dice Score (VoxelMorph) 0.8932;
Hausdorff Distance (affine registration) 6.557; Hausdorff Distance (VoxelMorph) 5.477

In the 2D atlas-based registration, the mean Dice score resulting from affine registration increases significantly after proper non-linear alignment, using VoxelMorph or SyN (ANTs). The decrease of the Hausdorff distances also reflects the improvement in alignment after deformable registration. Similar improvement is also observed in the case of 2D pairwise registration. In the experiment on 2D atlas-based registration, the Dice scores and the Hausdorff distances are almost similar for VoxelMorph and ANTs (SyN), with slightly better performance for the SyN method. In the experiment on 2D pairwise registration, similar conclusions are obtained, with in that case slightly better performance for VoxelMorph with respect to the Hausdorff distance.. For both experiments, VoxelMorph only takes 0.35 seconds on CPU to perform a single registration, in comparison to 15.19 seconds taken by ANTs (SyN), which represents a drastic reduction of the computational burden.

Similar conclusion can be drawn on a 3D pairwise registration experiment:VoxelMorph exhibits drastically reduced computation time, while achieving almost similar registration accuracy. It is also interesting to notice that VoxelMorph still performs well on data from another dataset (3D-IRCADb-01), thus illustrating the generalizability of the learned model. Here Dice scores and Hausdorff distance values are calculated not only for the liver but for five other segmented anatomical regions. In most cases, the performance obtained with VoxelMorph is similar to ANTs (SyN), while computational cost for a single registration estimation is less than 7 seconds on a single CPU for VoxelMorph compared to about 5 mins for ANTs. Concerning the influence of the similarity metric for the training of VoxelMorph, MSE and CC lead to similar performance, with a slight gain of performance with MSE. The drop in values of the Dice scores

for CC could be the effect of inconsistent anatomical structures present in the 3D volumes. While MSE is computed globally for the entire volumes, CC is calculated locally in our experiment. As the anatomical structures vary abruptly for the abdominal volumes, the performance of local CC for the entire abdominal volumes might not be up to the mark (although it is one of the best metrics). Hence we observed a slight decline in the Dice scores.

Fig. 6.8 and Fig. 6.9 present triplanar and 3D view of abdominal volume registration between two random test volumes taken from LiTS dataset and 3D-IRCADb-01 dataset respectively . The segmentations of the liver are superimposed on the original volumes. The visual inspection of the 3D views of registered liver volumes in Fig. 6.8 and Fig. 6.9 highlight that VoxelMorph and ANTs(SyN) lead to similar registration results, even when considering images from the 3D-IRCADb-01 dataset, which was not considered in the training set of VoxelMorph. This conclusion is also supported by the Dice scores and the Hausdorff distances.

Throughout the experiments, while visually inspecting the registration results on the test images, rare cases of poor registration were observed (samples are shown in Fig. 6.11 and Fig. 6.12). Although the training dataset of abdominal images (both 2D and 3D) was chosen and augmented carefully, it is clearly not possible to cover all anatomical configurations and variabilities on such a learning limited dataset. While it is not always feasible, it has been shown that providing anatomical segmentations along with the dataset for training can help to learn a model that better generalizes.

In this study we have assessed the quality of the registration using the standard Dice score, computed from the registration of segmented anatomical regions. If Dice is able to

evaluate the global correspondence of the organs, it does not provide information on the quality of the local matching. This limitation has been partly overcome by computing Dice scores on narrow structures such as the portal vein, whose Dice scores are sensitive to misregistrations. We have also provided Hausdorff distances which gives supplementary information. However, considering landmark-based evaluation metrics, such as "target registration errors" would definitely strengthen the evaluation.

6.5. Conclusion

This paper has shown that unsupervised learning-based approaches for 3D deformable registration become more and more competitive with conventional approaches. VoxelMorph has made a striking difference by substituting a time-consuming optimization problem for every test image pair with a learning-based registration algorithm, where a registration function is optimized over a training dataset. Precisely, after training, the model is able to perform deformable registration with state-of-the-art accuracy on abdominal images, while reducing the computation CPU time from minutes to seconds in comparison to SyN (ANTs). This paves the way for the routine use of deformable registration in the clinical workflow and will facilitate the development of high throughput image processing services.

References

1. Kim KW, Lee JM, Choi BI. Assessment of the treatment response of HCC. Abdominal imaging. 2011 Jun 1;36(3):300-14.

2. Christensen GE, Johnson HJ. Consistent image registration. IEEE transactions on medical imaging. 2001 Jul;20(7):568-82.

3. Klein A, Andersson J, Ardekani BA, Ashburner J, Avants B, Chiang MC, Christensen GE, Collins DL, Gee J, Hellier P, Song JH. Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. Neuroimage. 2009 Jul 1;46(3):786-802.

4. Avants BB, Tustison NJ, Song G, Cook PA, Klein A, Gee JC. A reproducible evaluation of ANTs similarity metric performance in brain image registration. Neuroimage. 2011 Feb 1;54(3):2033-44.

5. Balakrishnan G, Zhao A, Sabuncu MR, Guttag J, Dalca AV. An unsupervised learning model for deformable medical image registration. IEEE conference on Computer Vision and Pattern Recognition, 2018, pp. 9252-9260.

6. LiTS – Liver Tumor Segmentation Challenge (LiTS 2017); website: https://competitions.codalab.org/competitions/17094

7. 3D-IRCADb-01 database; website: https://www.ircad.fr/research/3d-ircadb-01/

8. Rohlfing T. Image similarity and tissue overlaps as surrogates for image registration accuracy: widely used but unreliable. IEEE Trans Med Imaging. 2012;31(2):153–163.

9. Crum WR, Hartkens T, Hill DL. Non-rigid image registration: theory and practice. The British journal of radiology. 2004 Dec;77(suppl_2):S140-53.

10. Sotiras A, Davatzikos C, Paragios N. Deformable medical image registration: A survey. IEEE transactions on medical imaging. 2013 Jul;32(7):1153.

11. Viola P, Wells III WM. Alignment by maximization of mutual information. International journal of computer vision. 1997 Sep 1;24(2):137-54.

12. Avants BB, Epstein CL, Grossman M, Gee JC. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Medical image analysis. 2008 Feb 1;12(1):26-41.

13. Dalca AV, Balakrishnan G, Guttag J, Sabuncu MR. Unsupervised learning for fast probabilistic diffeomorphic registration. International Conference on Medical Image Computing and Computer-Assisted Intervention 2018 Sep 16, pp. 729-738.

14. Rueckert D, Sonoda LI, Hayes C, Hill DL, Leach MO, Hawkes DJ. Nonrigid registration using free-form deformations: application to breast MR images. IEEE transactions on medical imaging. 1999 Aug;18(8):712-21.

15. Rohé MM, Datar M, Heimann T, Sermesant M, Pennec X. SVF-Net: Learning deformable image registration using shape matching. In International Conference on Medical Image Computing and Computer-Assisted Intervention 2017 Sep 10, pp. 266-274.

16. Sokooti H, de Vos B, Berendsen F, Lelieveldt BP, Išgum I, Staring M. Nonrigid image registration using multi-scale 3D convolutional neural networks. In International Conference on Medical Image Computing and Computer-Assisted Intervention 2017 Sep 10, pp. 232-239.

17. Yang X, Kwitt R, Styner M, Niethammer M. Quicksilver: Fast predictive image registration–a deep learning approach. NeuroImage. 2017 Sep 1;158:378-96.

18. Ronneberger O, Fischer P, Brox TN. Convolutional networks for biomedical image segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention 2015.

19. Eppenhof KA, Pluim JP. Pulmonary CT registration through supervised learning with convolutional neural networks. IEEE transactions on medical imaging. 2018 Oct 26;38(5):1097-105.

20. Murphy K, Van Ginneken B, Reinhardt JM, Kabus S, Ding K, Deng X, Cao K, Du K, Christensen GE, Garcia V, Vercauteren T. Evaluation of registration methods on thoracic CT: the EMPIRE10 challenge. IEEE transactions on medical imaging. 2011 May 31;30(11):1901-20.