Supervised learning in medical image registration

Sokooti, H.

Supervised Learning in Medical Image Registration

Hessam Sokooti

Colophon

About the cover:

The cover was designed from a symbolic transverse view of a lung and sagittal view of human faces. The checkerboard represents the displacement field. The design symbolically represents supervised registration with misalignment.

Supervised Learning in Medical Image Registration
Hessam Sokooti

ISBN: 978-94-92597-83-0
Thesis layout & cover designed by Hessam Sokooti
Printed by Boekendeal

Supervised Learning in Medical Image Registration

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof.dr.ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op donderdag 25 november 2021
klokke 16:15 uur

door

Hessam Sokooti

Promotor: Prof. dr. ir. B. P. F. Lelieveldt

Co-promotor: dr. ir. M. Staring

Leden promotiecommissie: Prof. dr. J. A Schnabel
Helmholtz Center Munich, Technical University Munich,
King's College London
Prof. dr. J. Batenburg
Prof. dr. M. Spruit
Prof. dr. S.A.R.B. Rombouts

Advanced School for Computing and Imaging

The research in this thesis was performed at the Division of Image Processing (LKEB), Department of Radiology of Leiden University Medical Center, The Netherlands. This thesis is financed by the Netherlands Organization for Scientific Research (NWO), project 13351.
This work was carried out in the ASCI graduate school. ASCI dissertation series number: 429.

Financial support for the publication of this thesis was kindly provided by:
ASCI research school,
Medis Medical Imaging,
Library of Leiden University
Bonitus Stichting,

Contents

1

Introduction

1.1 Medical image registration

Image registration is the process of aligning images by finding the spatial relation between the images. Assuming two images called fixed and moving images are taken at different time, different spatial location, or via a different imaging technique, the aim of image registration is to find an optimal transformation that aligns the fixed and the moving images.

Image registration has many applications in medical image analysis [1]. By aligning images from different modalities, the information can be fused together and provide complementary insight to a medical expert. For instance, in head and neck radiation therapy, Magnetic Resonance Imaging (MRI) provides higher contrast in soft-tissue. However, Computerized Tomography (CT) images commonly have better spatial resolution and provide electron density information. By aligning the MR and CT images, it is possible to exploit the advantage of each modality [2]. In fundus photography, image registration is also utilized to compose several images taken from different angles generating a single image with larger field of view (FOV). Fig. 1.1 illustrates an example of registration in fluorescein angiography (FA) retinal images [3]. Another application of the registration is in Alzheimer's disease classification with Brain MR images. The Jacobian of the transformation indicates the local volume change within the brain, which is an informative feature to detect Alzheimer's disease [4].

Given a fixed $\mathrm{I_F}(\boldsymbol{x})$ and a moving image $\mathrm{I_M}(\boldsymbol{x})$, the aim of a pair-wise registration is to find a displacement $u(\boldsymbol{x})$ that makes $\mathrm{I_M}\big(\boldsymbol{x}+u(\boldsymbol{x})\big)$ spatially aligned to $\mathrm{I_F}(\boldsymbol{x})$. The *transformation* is usually referred to as $\boldsymbol{T}(x)=\boldsymbol{x}+u(\boldsymbol{x})$. Here, we defined the direction of this mapping from the fixed image to the moving image as illustrated in Figure 1.1. In *parametric* registration, the transformation is defined by a model, like thin-plate

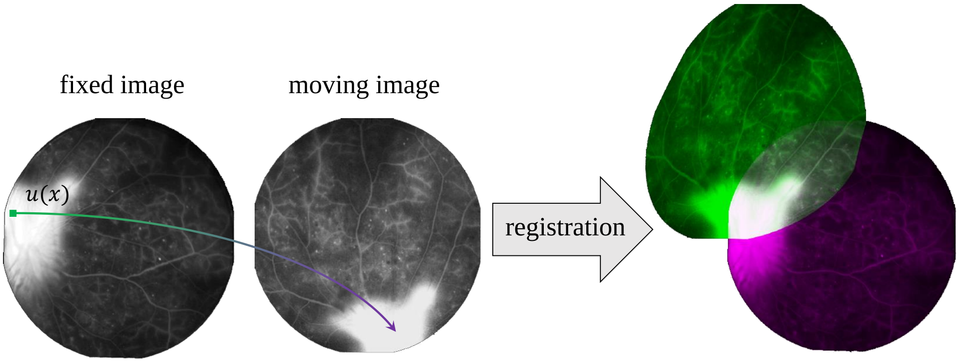


Figure 1.1: An example of two-dimensional image registration applied to fluorescein angiography retinal images. By aligning images taken from different angles, a single image with larger field of view (FOV) is composed. [3].

splines [5] or B-splines [6], while in *non-parametric* registration, the degree of freedom of the transformation is equal to the total number of voxels in the image. After finding the transformation, the moving image will be resampled in the fixed image domain by an interpolation technique.

Conventionally, image registration is cast to an optimization problem. A loss function is defined based on a dissimilarity measure. For instance, a simple loss can be defined as a mean squared difference (MSD) of the intensity values of the overlapping region between the fixed and the moving images. This optimization can be solved with an iterative approach like stochastic gradient descent [7]. Finally, the optimal transformation will be found with respect to the loss function.

Fine-tuning an image registration algorithm is a time-consuming task. Both the dissimilarity metric and the transformation model need to be selected and tuned in order to achieve high quality registration performance. Another drawback of conventional image registrations is that their inference time is rather slow. Since most of them use an iterative optimization method, it is not trivial to run the optimization in parallel. Fast image registration is required in several medical tasks such as registering the follow-up scans in adaptive radiotherapy [8] and image-guided surgery [9].

1.2 Learning-based image registration

Learning-based registration techniques are becoming more popular [10]. One of the applications is to learn a dissimilarity metric instead of tuning over handcrafted dissimilarity metrics. The advantage of learning a dissimilarity metric is more prominent in multi-modal image registration like ultrasound (US)/MR, [11, 12]. It is reported that the learned metric outperforms well-known multimodal metrics such as mutual information (MI) [13] and the modality independent neighbourhood

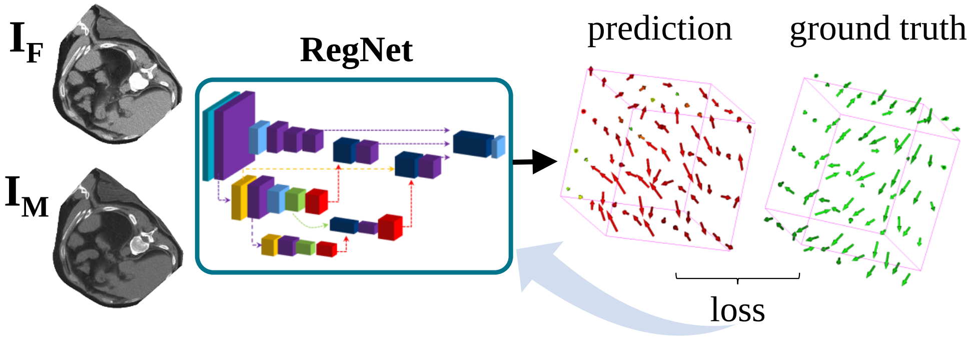
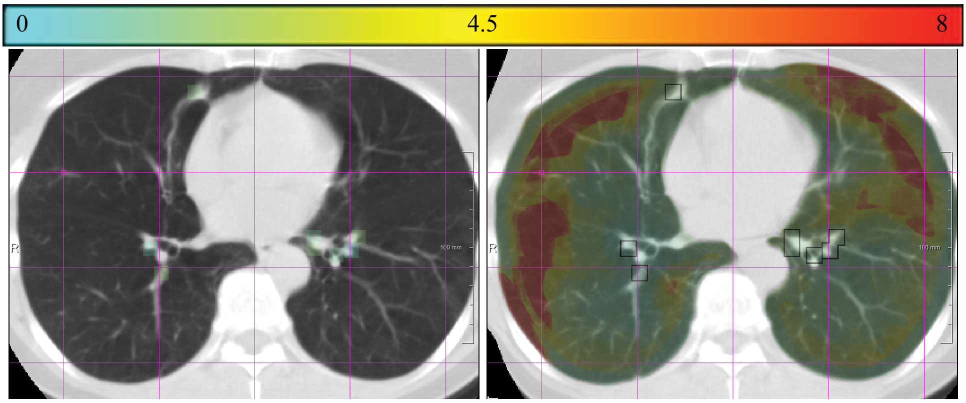


Figure 1.2: A schematic view of convolutional neural network based registration with the supervised transformation approach. The loss function is computed based on the dissimilarity between a ground truth transformation and the predicted one from the network.

descriptor (MIND) [14].

Several methods have been proposed to learn the entire image registration pipeline using deep learning (DL). Most methods were introduced after starting this PhD theses. In reinforcement learning (RL) approaches [15], instead of a conventional optimization, a trained agent is used to perform the registration [16]. However, the RL approach can still be time-consuming. In principle, both optimization and dissimilarity metrics can be learned simultaneously. Thus, at the inference time, the registration is usually performed in one shot. In the supervised transformation approach, a known transformation is used during the training [17, 18, 19, 20, 21]. The loss function then can be defined based on the difference between the the known transformation and the predicted transformation. Fig. 1.2 illustrates a schematic design of this approach. To train the network, three inputs are needed, the fixed image, the moving image and the known ground truth transformation between the fixed and the deformed moving image. In the unsupervised transformation approach, an indirect loss is utilized to guide the transformation. Several examples of this indirect loss are the mutual information, cross correlation ([22, 23, 24]), the Dice overlap of known segmentation maps [25], normalized gradient field distance measure [26], or even using generative adversarial networks (GAN [27]) to learn a new indirect loss [8, 28]. Recent papers showed that utilizing a proper regularization such as bending energy [18], volume change penalty [26] or a graph regularization network on a keypoint-based registration [29] can improve the performance of registration.



(a) Fixed Image (b) Deformed moving image

Figure 1.3: An example of registration error map, which is overlaid on the deformed moving image in a chest CT scan pair [33]. The color bar indicates the estimated registration error in mm.

1.3 Uncertainty and error in image registration

In most registration methods, no assessment of the registration quality is provided, and simply the result is returned. Many medical pipelines are based on registered images and it is important to know the uncertainty of registration before continuing to the next phase in order to prevent the accumulation of errors. For example, in online adaptive radiotherapy daily contouring of the tumor and organs-at-risk can be performed with the help of image registration and therefore quality assessment (QA) is mandatory to ensure patient safety [30]. Visualizing the registration error can also be directly helpful in various medical applications. For instance, an error map of registering a pre-operative scan and an atlas could provide more insight about the localization error during a surgical procedure [31]. Refinement of registration is another important application of error prediction [32]. An example of a registration error map is given if Fig. 1.3. The color bar indicates the estimated registration error in mm. For instance, the registration probably should be improved in the red regions. This error map provide insight about the registration quality. Hence, a medical expert can consider the local uncertainty of the subsequent analysis on the aligned images. Currently, the quality assessment of registration is usually performed manually, which is a time-consuming task and prone to human fatigue.

Defining the registration error is not trivial as image registration is an ill-posed problem. The registration error can be better explained on the corresponding distinctive landmark locations. However, computing the registration error over homogeneous regions is more challenging. Registration uncertainty can be counted as a measure of confidence in the registration output. Probabilistic image registration (PIR) methods usually can provide transformation and uncertainty at once [34]. In registration methods using continuous optimization, one way to estimate the uncertainty is to perturb the initial state [35]. The uncertainty sometimes is used as a surrogate for registration error. It should be noted that high uncertainty does not always means high registration error and vice versa [36].

Several naive intensity-based and registration-based features were proposed as a surrogate for registration misalignment, such as local normalized mutual information (NMI) [37] and the local gradient of the loss function [38]. Simply, the smaller value of NMI or the gradient indicates smaller registration misalignment. More advanced learning techniques are also utilized in predicting the registration error such as learning over landmark locations [39] and learning over artificially generated image pairs [40]. However, the accuracy of naive methods are not promising and the inference of the advanced methods is time-consuming.

1.4 Outline of the thesis

The aim of this thesis is to develop a learning-based image registration method as a much faster alternative to conventional methods without requiring hyper-parameter tuning. We also aimed to improve accuracy as well as inference time of registration misalignment detection methods, via a fully automatic solution. Although all the proposed methods in this thesis are generic, all the experiments are performed on chest CT scans.

Chapter 2 presents a novel method to solve nonrigid image registration through a learning approach, instead of via iterative optimization of a predefined dissimilarity metric. We design a Convolutional Neural Network (CNN) architecture that, in contrast to all other work, directly estimates the displacement vector field (DVF) from a pair of input images. This is one of the first methods proposed in literature to solve nonrigid image registration via deep learning.

Chapter 3 extends chapter 2 into a practical pipeline based on efficient supervised learning from artificial deformations. The proposed architectures are embedded in a multi-stage approach to increase the capture range of the networks in order to more accurately predict larger displacements.The proposed method achieved the best result on the DIR-Lab 4DCT study among all published DL-based registration methods up to the publication date.

Chapter 4 proposes a new automatic method to predict the registration error in a quantitative manner and is applied to chest CT scans. A random regression forest is utilized to predict the registration error locally. The forest is built with features related to the transformation model and features related to the dissimilarity after registration. Several of the proposed features are novel and unique as well.

Chapter 5 presents a supervised method to predict registration misalignment using convolutional neural networks (CNNs). This task is casted to a classification problem with multiple classes of misalignment: “correct” 0-3 mm, “poor” 3-6 mm and “wrong” over 6 mm. Rather than a direct prediction, we propose a hierarchical approach, where the prediction is gradually refined from coarse to fine. Our solution is based on a convolutional Long Short-Term Memory (LSTM), using hierarchical misalignment predictions on three resolutions of the image pair, leveraging the intrinsic strengths of an LSTM for this problem.

Chapter 6 summarizes and discusses the overall achievements of this thesis.

2

Nonrigid Image Registration using Multi-Scale 3D Convolutional Neural Networks

This chapter was adapted from:

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

Abstract

In this chapter, we propose a method to solve nonrigid image registration through a learning approach, instead of via iterative optimization of a predefined dissimilarity metric. We design a Convolutional Neural Network (CNN) architecture that, in contrast to all other work, directly estimates the displacement vector field (DVF) from a pair of input images. The proposed RegNet is trained using a large set of artificially generated DVFs, does not explicitly define a dissimilarity metric, and integrates image content at multiple scales to equip the network with contextual information. At testing time nonrigid registration is performed in a single shot, in contrast to current iterative methods. We tested RegNet on 3D chest CT follow-up data. The results show that the accuracy of RegNet is on par with a conventional B-spline registration, for anatomy within the capture range. Training RegNet with artificially generated DVFs is therefore a promising approach for obtaining good results on real clinical data, thereby greatly simplifying the training problem. Deformable image registration can therefore be successfully casted as a learning problem.

2.1 Introduction

Deformable image registration (DIR) is the task of finding the spatial relationship between two or more images, and is abundantly used in medical image analysis. Typically, image registration is solved by iteratively optimizing a predefined handcrafted intensity-based dissimilarity metric over the transformation parameters. The metric represents a model of the intensities encountered in the image data. Problems may occur when part of the data does not fit the model, which are typically dealt with by making modifications to the dissimilarity metric. Instead, in this paper we take another approach, where we do not handcraft such a model, but use a machine learning approach to automatically determine what constitutes an accurate registration, i.e. without explicitly defining a dissimilarity metric. The proposed method is based on regression using Convolutional Neural Networks (CNNs), that directly learns a displacement vector field (DVF) from a pair of input images.

The idea of learning registration has shown to be promising [41]. Several CNN regression techniques have been proposed in the context of image registration. Miao *et al.* [42] applied CNN regression for rigid 2D-3D registration. Liao *et al.* [16] used CNN regression to model a sequence of motion actions for 3D registration. Their method is iterative (not one shot), and limited to rigid-body transformations. For nonrigid approaches, Yang *et al.* [43] predicted the initial momentum of a 3D LDDMM registration. Eppenhof *et al.* [44] trained a CNN to predict the local registration error, without performing a full registration. Related work has been done in the field of optical flow [45].

In contrast, we propose an end-to-end method that directly predicts the 3D nonrigid DVF given a fixed and a moving image, without requiring a dissimilarity metric like conventional methods. The proposed architecture, called RegNet, analyzes 3D input patches at multiple scales to equip the CNN with contextual information. Training is based on a wide variety of artificial displacements acting as the target value in the loss function, while testing is performed on registration of baseline and follow-up CT images of a patient. At testing time the registration is performed in a single shot, in contrast to current iterative methods. To the best of our knowledge this is the first method that solves nonrigid 3D image registration with CNNs end-to-end, i.e. directly predicting DVFs.

2.2 Methods

2.2.1 Network architecture

The proposed CNN architecture RegNet takes patches from a pair of 3D images (the fixed image I_F and the moving image I_M) as input. The output of RegNet is a vector of three elements, which is the displacement of the central voxel of the patch. A full DVF

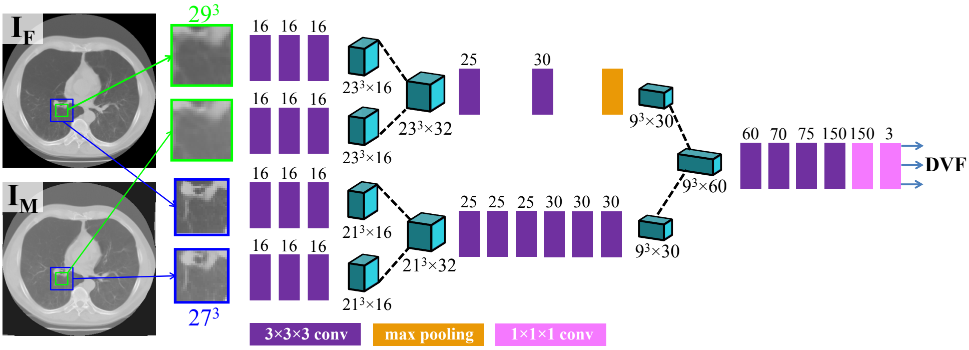


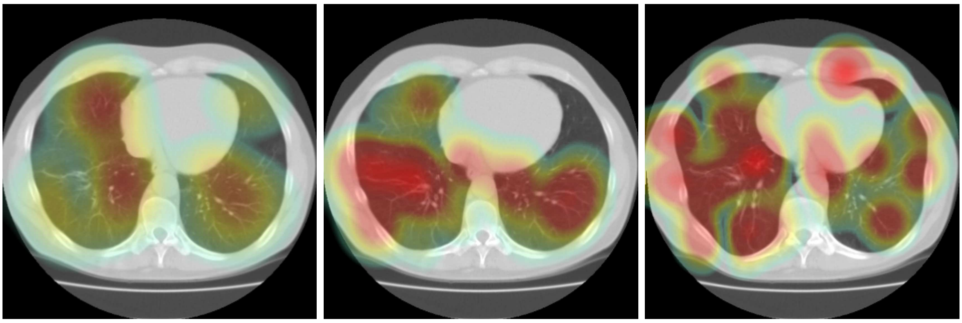
Figure 2.1: RegNet design.

is generated by sliding over the input images. The DVF is defined as the displacement $\boldsymbol{u}(\boldsymbol{x})$, mapping points from the fixed image domain to that of the moving image. The transformation is defined as $\boldsymbol{T}(\boldsymbol{x}) = \boldsymbol{x} + \boldsymbol{u}(\boldsymbol{x})$.

For each image we extract patches at original resolution of size 29x29x29 voxels. To improve the receptive field of the network, we additionally extract patches of 54x54x54 voxels, which are downsampled to 27x27x27 voxels. In this way local as well as more global information is incorporated, allowing better discrimination between anatomical locations and to add contextual information. The downsampling makes sure there is limited effect on memory consumption and computational overhead. Similar multi-scale approaches have been shown effective for segmentation [46]. We thus have four 3D patches as inputs.

We start with three convolutional layers for each input patch separately (late-fusion) instead of stacking them as channels (early-fusion). The fixed and moving patches of each resolution are then merged by concatenation. This is followed by 2 and 6 convolutional layers for the original resolution and the downsampled patch, respectively. Max pooling is used on the pipeline of the original resolution, ensuring spatial correspondence of the activation of the two pipelines before merging; for every 2 shift of the receptive field of the original resolution only 1 shift should be performed in the low resolution [46]. The two resolution pipelines are then also concatenated, followed by 4 convolutional layers and two fully connected layers. All convolutional layers use $3 \times 3 \times 3$ kernels, batch normalization and ReLu activation. The network architecture is visualized in Fig. 2.1.

Optimization is done using Adam, with a decaying learning rate starting at 0.001 and a decay factor of 1.25 in each epoch, which improved the convergence rate in our experiments. The loss function is defined as the mean residual distance between target and estimated DVF: $\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\mathrm{DVF}'_i - \mathrm{DVF}_i|$, with DVF′ the prediction of RegNet



(a) Low frequency (b) Medium frequency (c) High frequency

Figure 2.2: Heat maps of the magnitude of DVFs used for training RegNet.

and DVF the target defined in Section 2.2.2.

2.2.2 Training

To train our network, synthetic DVFs are generated with varying spatial frequency and amplitude, aiming to represent the range of displacements that can be seen in real images: 1) Creating a vector field with the size of the input image (which will act as the moving image) and initializing it with zero vectors; 2) Randomly selecting P points in the DVF and randomly assigning three values to the displacement vector in the range $[-\theta, +\theta]$; 3) Smoothing the DVF with a Gaussian kernel with a standard deviation of σ. Low, medium and high frequency deformations are generated using the settings $\sigma = 35, P = 80, \theta = 8$; $\sigma = 25, P = 100, \theta = 8$; and $\sigma = 20, P = 100, \theta = 8$, respectively. Transformed images are generated by applying the DVF to the input image, using cubic B-spline interpolation, resulting in the fixed image. To allow more accurate simulation of real images, Gaussian noise with a standard deviation of 5 is finally added to the images. Examples are available in Fig. 2.2.

It is possible to generate plenty of deformations for a single moving image, but a drawback of this approach is that the moving image is identical in each pair of input images, as only the fixed image is generated randomly. We therefore also generate deformed versions of the moving image, based on which new deformed images are created. The new moving images are generated using low frequency deformations only, to avoid over-stretching (leading to a blurred appearance). We use the settings $\sigma = 35, P = 100, \theta = 8$ and Gaussian noise with a standard deviation of 3 in this step.

2.3 Experiments and results

2.3.1 Materials

We use data from the SPREAD study [47], which contains 19 pairs of 3D chest CT images. The dimension of the images is about $446 \times 315 \times 129$ with an average voxel size of $0.781 \times 0.781 \times 2.5$ mm. Patients are between 49 and 78 years old and for each patient a baseline image and a 30 months follow-up image are available. For each pair, 100 well-distributed corresponding landmarks were previously selected semi-automatically at distinctive locations [48]. All images were resampled to a voxel size of $1 \times 1 \times 1$ mm.

RegNet is written in Theano [49] and Lasagne [50], artificial DVFs are created using SimpleITK [51]. Conventional registrations are performed using `elastix` [52].

2.3.2 Experimental setup and evaluation

The set of 19 image pairs is divided in a training set of 10 pairs, a validation set of 2 pairs, and a test set of 7 pairs. 2100 patches per image are randomly extracted from the lung regions of the training images, using both the baseline and follow-up images as input for training. For each image in the database we create 6 different DVFs (3 for a single moving image and 3 other after deforming that moving image, see Section 2.2.2), resulting in $252,000$ training examples. In addition, we applied data augmentation, flipping all patches in the x, y and z direction and adding Gaussian noise with a standard deviation of 5. In total we have approximately 1 million patches available for training. The network is trained for 15 epochs. The validation set was used to monitor overfitting during training, and to compare with the single-scale and the early-fusion design.

The test set was used in two ways. We first evaluate the ability of the trained network to register artificially deformed image pairs, which is how RegNet was trained. This was evaluated using the MAE measure. Second, we apply RegNet for registration of the real baseline and follow-up CT images, without artificial deformations. This experiment is evaluated using the set of corresponding landmarks, where we report their mean Euclidean distance after registration: $\mathrm{TRE} = \frac{1}{n}\sum_{i=1}^{n} \|\mathrm{DVF}'_i(\boldsymbol{x}_F) + \boldsymbol{x}_F - \boldsymbol{x}_M\|_2$, with $\boldsymbol{x}_F$ and $\boldsymbol{x}_M$ the landmark locations. An initial affine registration is performed before applying RegNet, similar to conventional approaches. We use an intensity-based method (normalized correlation), using 5 resolutions of 1000 iterations each. RegNet is compared with two conventional B-spline registrations with a final grid spacing of 10mm: a version using a single resolution of 2000 iterations, and one using 3 resolutions of 500 iterations each. As the capture range of our network is certainly less than half the patch width, we additionally present the TRE of only those points that are within 8mm distance after the affine registration (TRE').

2.3.3 Results

All quantitative results are given in Table 2.1. The results on the validation set show that multi-scale late-fusion RegNet performs better than either single-scale or early-fusion RegNet. It can be seen that the regression accuracy on the validation set (MAE) is about 1 mm, showing that RegNet was successfully trained. The results in the x and

Table 2.1: Quantitative results

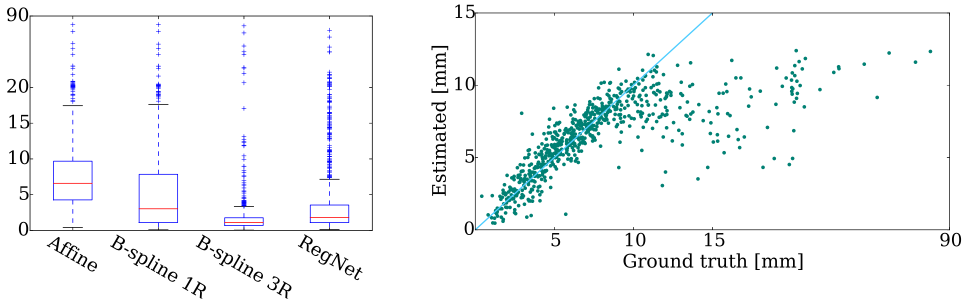
Evaluation	Method	Data	measure	measure$_x$	measure$_y$	measure$_z$
MAE	RegNet 1Scale	validation	1.70 ± 1.81	0.56 ± 0.78	0.53 ± 0.71	0.61 ± 0.88
	RegNet Early	validation	1.26 ± 1.22	0.41 ± 0.51	0.39 ± 0.48	0.45 ± 0.60
	RegNet	validation	1.17 ± 1.10	0.36 ± 0.56	0.38 ± 0.44	0.43 ± 0.49
	RegNet	test	1.19 ± 1.17	0.36 ± 0.59	0.40 ± 0.50	0.43 ± 0.51
TRE	Affine	test	8.08 ± 7.18	4.21 ± 4.40	3.92 ± 5.64	3.80 ± 4.25
	B-spline 1R	test	5.48 ± 7.56	2.47 ± 4.01	2.64 ± 5.71	2.92 ± 4.12
	B-spline 3R	test	2.19 ± 6.22	0.67 ± 1.97	1.04 ± 5.07	1.45 ± 3.21
	RegNet	test	4.39 ± 7.54	2.19 ± 4.53	1.79 ± 4.83	2.35 ± 4.33
TRE′	Affine	test	5.39 ± 2.25	2.80 ± 2.04	2.70 ± 1.92	2.73 ± 1.93
	B-spline 1R	test	2.59 ± 2.28	1.02 ± 1.44	1.09 ± 1.47	1.72 ± 1.56
	B-spline 3R	test	1.28 ± 0.94	0.41 ± 0.51	0.42 ± 0.43	1.00 ± 0.86
	RegNet	test	1.66 ± 1.26	0.58 ± 0.62	0.64 ± 0.77	1.19 ± 1.10

y direction are slightly better than that in the z direction, which can be attributed to the relatively large slice thickness of our data. The MAE results on the test set confirm that RegNet can successfully register artificially deformed images with a sub-voxel accuracy.

For the test set we have 685 corresponding landmarks available to compute the TRE. For TRE′, 503 landmarks are within 8 mm after affine registration. The results for affine, the two B-spline settings and RegNet are listed in Table 2.1 and illustrated in Figs. 2.3 and 2.4. It can be seen that the multi-resolution B-spline method overall gives the best performance (TRE results), but RegNet is better than a single resolution B-spline. When we focus on the points within the capture range of RegNet (TRE′ results) it can be seen that RegNet performs better than the single resolution B-spline method, and performs similar to multi-resolution B-spline. For those landmarks a residual error of 1.7 mm is obtained, which is sub-voxel with respect to the original resolution. Again, the accuracy in the x and y direction is slightly better than that in the z direction. Fig. 2.3b shows a scatter plot of all landmarks after registration with RegNet. RegNet gives accurate registrations until ~8 mm, which is to be expected due to the patch size and the fact that RegNet was trained up to $\theta = 8$ mm deformations only. Figs. 2.4b-d show scatter plots of the landmarks within 8 mm, for the three directions separately. Example registration results are given in Fig. 2.5. Inference time for an image of size 300^3 is about 14 seconds on a Tesla K40.

2.4 Discussion and conclusion

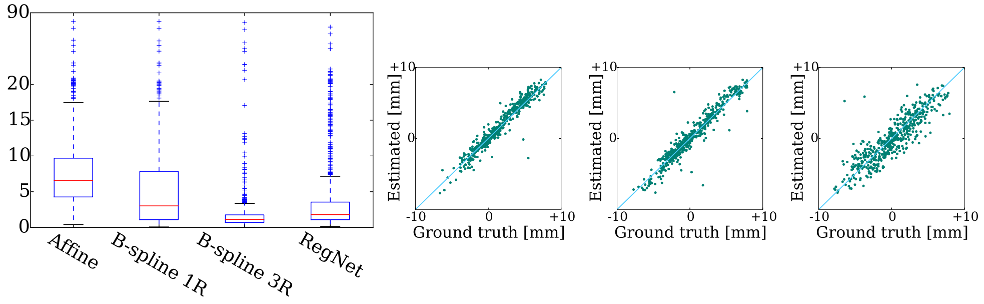
We presented a convolutional neural network (RegNet) for 3D nonrigid image registration. RegNet can be successfully applied to real world data, after training on artificially generated displacement vector fields. Tests on artificially deformed images



(a) box-plot

(b) Scatter plot for RegNet

Figure 2.3: Residual landmark distances, for all landmarks.



(a) box-plot

(b) x direction

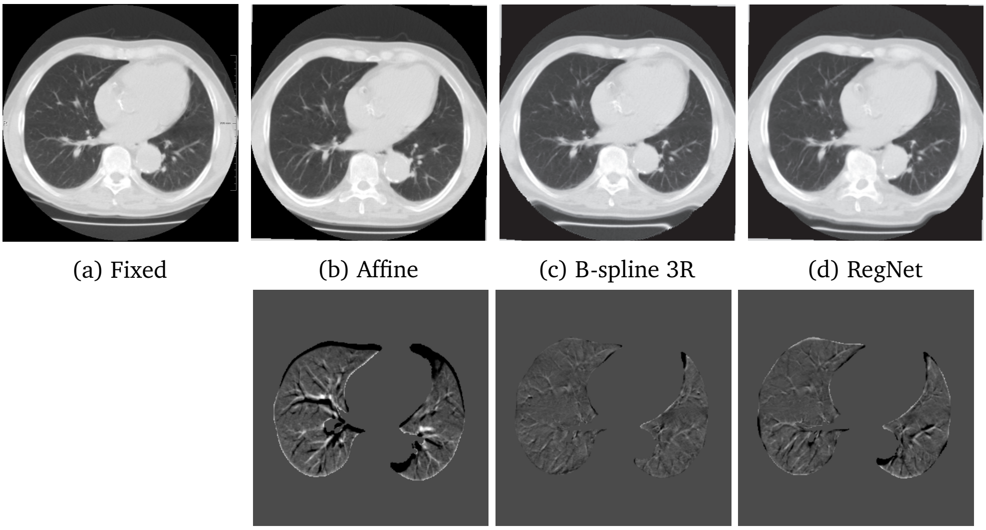
(c) y direction

(d) z direction

Figure 2.4: Residual landmark distances, for the landmarks in the capture range. Figures b-d show scatter plots of RegNet against the ground truth.

as well as with intra-patient chest CT data, showed that RegNet achieved sub-voxel registration performance, for landmarks within the capture range. This was better than the performance of a conventional single resolution B-spline registration method, and close to that of a multi-resolution B-spline. When considering all landmarks, the multi-resolution B-spline method still outperformed RegNet. In the training phase of RegNet no use was made of (manually annotated) corresponding points, or segmentations for guidance, which are hard to obtain in large quantities. Synthetic DVFs on the other hand can easily be generated in bulk, which greatly simplifies the training process.

In our current design the registration capture range is related to the size of the patches that are shown to the network, and the results show good performance until 8 mm, but deteriorate after that. The capture range may be enlarged by the use of larger patches or the addition of more scales to the network. It is also possible to extend RegNet to a multi-resolution approach, working from even further downsampled (and



(a) Fixed (b) Affine (c) B-spline 3R (d) RegNet

(f) Affine (g) B-spline 3R (h) RegNet

Figure 2.5: Example results (top row) and difference images (bottom row).

smoothed) images than in the current multi-scale approach, successively upsampling until the original resolution.

For future work, we will perform a sensitivity analysis of a number of important parameters of RegNet, like the patch size and its relation to the several parameters that define the training DVFs (e.g. the maximum magnitude θ). We will also train RegNet in other applications besides chest CT, to test the generalizability of the architecture.

In conclusion, the proposed neural network achieves promising results for the nonrigid registration of image pairs, using an end-to-end approach. Information at multiple scales is integrated in the CNN. After training, deformable registration is performed in one shot.

Acknowledgments

This work is financed by the Netherlands Organization for Scientific Research (NWO), project 13351. Dr. M.E. Bakker and J. Stolk are acknowledged for providing a ground truth for the SPREAD study data used in this paper. The Tesla K40 used for this research was donated by the NVIDIA Corporation.

3

3D Convolutional Neural Networks Image Registration Based on Efficient Supervised Learning from Artificial Deformations

This chapter was adapted from:

H Sokooti, B de Vos, F Berendsen, M Ghafoorian, S Yousefi, BP Lelieveldt, I Išgum, and M Staring. **3D Convolutional Neural Networks Image Registration Based on Efficient Supervised Learning from Artificial Deformations**, *arXiv preprint arXiv:1908.10235*, 2019.

Abstract

We propose a supervised nonrigid image registration method, trained using artificial displacement vector fields (DVF), for which we propose and compare three network architectures. The artificial DVFs allow training in a fully supervised and voxel-wise dense manner, but without the cost usually associated with the creation of densely labeled data. We propose a scheme to artificially generate DVFs, and for chest CT registration augment these with simulated respiratory motion. The proposed architectures are embedded in a multi-stage approach, to increase the capture range of the proposed networks in order to more accurately predict larger displacements. The proposed method, RegNet, is evaluated on multiple chest CT scans studies and achieved a target registration error of 2.32 ± 5.33 mm and 1.86 ± 2.12 mm on SPREAD and DIR-Lab-4DCT studies, respectively. The average inference time of RegNet with two stages is about 2.2 s.

3.1 Introduction

Image registration is the process of aligning images and has many applications in medical image analysis. Generally, image registration casts to an optimization problem of minimizing a predefined handcrafted intensity-based dissimilarity metric over a transformation model. Both the dissimilarity metric and the transformation model need to be selected and tuned in order to achieve high quality registration performance. This task is time-consuming and there is no guarantee that the selected dissimilarity model fits with new images.

General learning-based techniques have been used in several registration papers. Guetter et al. [53] incorporated a prior learned joint intensity distribution to perform a nonrigid registration. Jiang et al. [54] selected and fused a large number of features instead of using only one similarity metric. Hu et al. [41] leveraged regression forests to predict an initial DVF. In terms of predicting registration accuracy Muenzing et al. [39] casted this task to a classification problem and extracted several local intensity-based features, which are fed to a two-stage classifier. Sokooti et al. [55, 33] extracted some intensity-based and registration-based features, then by using regression forests estimated the local registration error.

In recent years, CNNs have also been utilized in the context of image registration. Miao et al. [42] used CNNs for rigid-body transformations. Yang et al. [43] trained a CNN to predict the initial momentum of a 3D LDDMM registration. Cao et al. [56] generated a multi-scale similarity map and utilized it to predict the DVF. Simonovsky et al. [57] proposed a CNN-based similarity metric for multi-modal registration. Their training samples were a set of aligned images as the positive cases and a set of manually deformed images as the negative cases.

In the unsupervised deep learning approaches, de Vos et al [22, 23] for the first time used normalized cross correlation (NCC) of the fixed and moving image as a loss function. Later Balakrishnan et al. [24] and Ferrante et al. [58] used the same loss to train their network. Mahapatra et al. [25] combined NCC with other similarity metrics such as the Dice overlap metric over the labeled images. Elmahdy et al. [8] utilized an adversarial training based on the segmentation maps in addition to the NCC loss. Sheikhjafari et al. [59] and Dalca et al. [60] employed the mean squared intensity difference, which was applied to mono-modal image registration. Hu et al. [61] proposed a loss function that calculates cross entropy over the smoothed segmentation maps, which was applied to multi-modal images. A drawback to use conventional similarity metrics is that these similarity metrics are not perfect and might not fit in all images.

In the supervised approaches, for the first time Sokooti et al. [17] generated artificial DVFs with different frequencies to train a CNN architecture. Rohé et al. [62]

proposed to build reference DVFs which were obtained by performing registration over segmented regions of interest. Fan et al. [63] proposed a ground truth based on the GAN network. The implicit ground truth is assigned using the negative cases derived from the generator network while the positive cases are synthetically made by perturbing the original images. Eppenhof et al. [19] constructed a small set of images by applying a random DVF. In the model-based methods Uzunova et al. [64] proposed statistical appearance models to be used for data augmentation. Hu et al. [65] utilized biomechanical simulations to regularize their network.

In several articles, reinforcement learning is used [66, 67, 68]. An artificial agent is trained by making a statistical deformation model from training data. However, this approach is still iterative and might be slow at inference time.

Conventionally, in quality assessment of registration, manually selected landmarks or manually segmented regions are used. However, utilizing them as a gold standard in training has some drawbacks. With manually segmented regions, several measurements like Dice and mean surface distance can be calculated, but there is no direct correlation between Dice and the true DVF in all voxels of the image. The drawback of using landmarks as a gold standard [55, 33] is that the numbers of landmarks usually is not enough to estimate a continuous gold standard DVF for the whole image.

In this paper, instead of using a transformation model, we directly predict the displacement vector field (DVF). The convolutional neural network (CNN) implicitly learns the dissimilarity metric. The current paper is a large extension of the work first presented in Sokooti et al. [17]. We present more ways to construct sufficiently realistic synthetic DVFs. The network design is greatly enhanced by increasing the capture range in order to more accurately predict larger DVFs. A multi-stage approach is also proposed to overcome this issue. The evaluation is performed on the SPREAD study as well as on the public DIR-Lab study. The proposed method is capable to be trained on any datasets without needing any manual ground truth.

3.2 Methods

3.2.1 System overview

A block diagram of the proposed system is given in Fig. 3.1. The inputs of the system are a fixed image I_F and a moving image I_M. Similar to the conventional registration methods, a multi-stage approach is employed. The registration blocks RegNet^4 and RegNet^2 perform on the down-scaled images with a factor of 4 and 2, respectively. The inputs of the final registration block RegNet^1 are original resolution images. The output of the system is a predicted DVF of transforming the moving image to the fixed image which is defined as $\boldsymbol{T}(\boldsymbol{x}) = \boldsymbol{T_{s1}}\big(\boldsymbol{T_{s2}}\big(\boldsymbol{T_{s4}}(\boldsymbol{x})\big)\big)$.

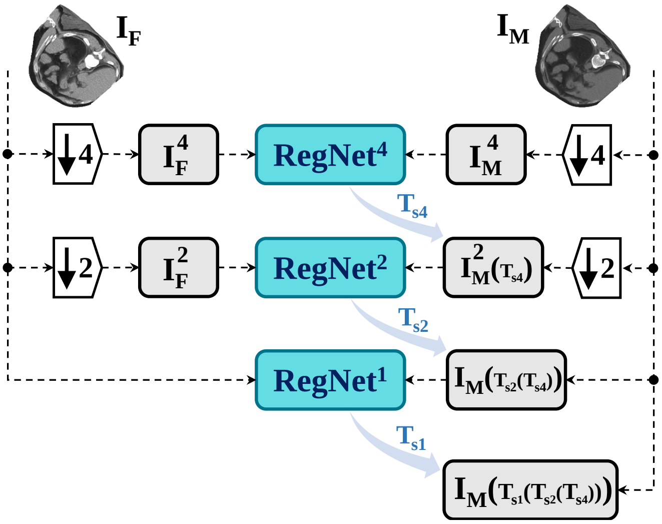


Figure 3.1: Block diagram of the proposed system. The initial inputs of the system are fixed and moving images down-scaled by a factor of four (↓4). Three RegNets process the input images over three stages (4, 2, 1) and generate the final output $\boldsymbol{T}(\boldsymbol{x}) = \boldsymbol{T}_{s1}\big(\boldsymbol{T}_{s2}\big(\boldsymbol{T}_{s4}(\boldsymbol{x})\big)\big)$.

3.2.2 Network architecture

We propose three network architectures for the RegNet design. The first two architectures are patch-based, and predict the DVF for a local neighborhood. These two networks are more complex and occupy a relatively large amount of GPU memory. The third architecture is based on a more simple U-Net design [69] with fewer network weights, and is capable of registering entire images (not patches), but down-scaled, within the memory limits of current GPUs. This last architecture is considered a candidate for the first resolution (RegNet^4), while the others are considered for the second and third resolution (RegNet^2 and RegNet^1). In Section 3.3 we compare these architectures and combinations thereof.

The networks have some settings in common. All convolutional layers use batch normalization [70] and ReLu activation [71], ,except for the last two layers of the U-Net design and the last three layers of the patch-based designs, where ELu activation is used to improve the regression accuracy. The last layer of all architectures does not use batch normalization nor an activation function. The Glorot uniform initializer [72] is used for all convolutional layers except for the trilinear upsampling, in which a fixed trilinear kernel is utilized. The three architectural designs are given in Fig. 2. The details are:

3.2.2.1 U-Net (U)

U-Net is one of the most common designs used in medial image segmentation The proposed modified design has an input size and output size of $125 \times 125 \times 125$ voxels. This architecture is only used for the sub-block RegNet^4, i.e. CNN-based registration is applied to down-scaled images with a factor of four. The proposed design is given in Fig. 3.2a. This relative simple design has 232,749 trainable parameters.

3.2.2.2 Multi-View (MV)

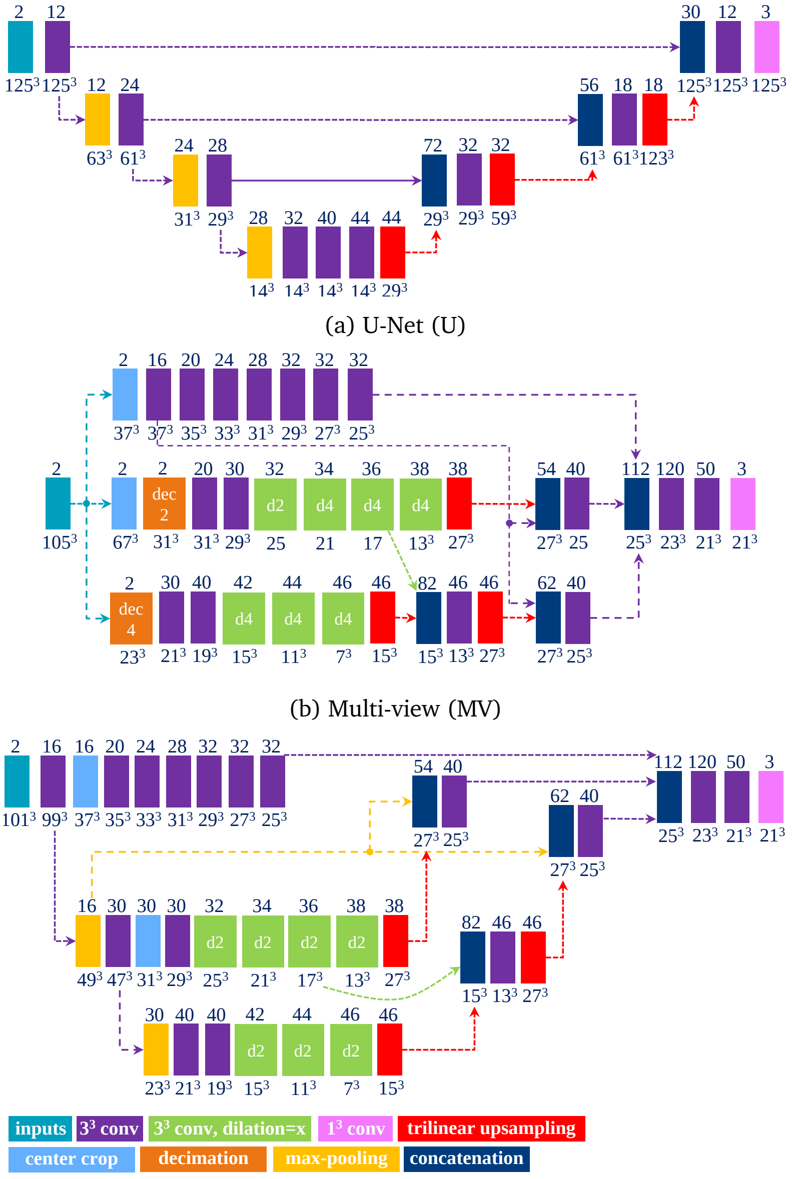
In this design, different scales are created by using conventional decimation by convolving the inputs with fixed B-spline kernels, which is similar to [17] and [46]. This design is relatively more memory efficient because of this multi-view approach. The proposed CNN architecture is visualized in Fig. 3.2b. The input of the network is a pair of 3D patches of size $105 \times 105 \times 105$ for the fixed and moving image. The network is then split into 3 pipelines: down-scaled with a factor of 4, a factor of 2, and the original resolution. In order to save memory, the original resolution and the down-scaled version with a factor 2 are cropped to $37 \times 37 \times 37$ and $67 \times 67 \times 67$, respectively. Decimation is done with the help of convolutions with a fixed B-spline kernel. In the down-scaled factor 2 pipeline, a stretched B-spline kernel with size $7 \times 7 \times 7$ is used. For down-scaling with a factor of 4, the B-spline kernel is stretched by a factor of 4, and has a size of $15 \times 15 \times 15$. Each pipeline continues with several convolutional layers with dilation of 1 or higher. The upsampling layers ensure spatial correspondence of all three pipelines. Finally, all pipelines are merged together followed by three more convolutional layers. The network gives three 3D outputs of size $21 \times 21 \times 21$ corresponding to the displacement in x, y and z direction. The total number of parameters in this design is 1,201,353.

3.2.2.3 U-Net-Advanced (Uadv)

This proposed architecture is again a patch-based one but using a max-pooling technique instead of a decimation method. The global design is similar to the U-net architecture, but instead of simple shortcut connections, several convolutions are used for these connections. The proposed design is illustrated in Fig. 3.2c. The network starts with a convolutional layer to extract several low-level features from the images before any max-pooling. The size of the inputs and output are $101 \times 101 \times 101$ and $21 \times 21 \times 21$. The total number of parameters in this design is 1,420,701.

3.2.3 Artificial generation of DVFs and images

In order to train a CNN, a considerable number of ground truth DVFs are needed. We take a moving image I_M from the training set. The fixed image I_F is created artificially by generating a DVF, applying the DVF to the moving image resulting in I_F^{clean}, and adding artificial intensity models to finally obtain I_F.



(a) U-Net (U)

(b) Multi-view (MV)

(c) U-Net-advanced (Uadv)

Figure 3.2: RegNet designs: The inputs of the U-Net design are entire down-scaled images. However, in the Multi-view and U-Net-advanced architectures the output size is smaller than the input size and can be trained in a patch-based manner.

3.2.3.1 Artificial DVF

We propose to generate three categories of DVFs, to represent the range of displacements that can be seen in real images:

single frequency: The first category consists of DVFs having one or more local displacements of only one spatial frequency. They are generated as follows: Create an empty B-spline grid of control points with a spacing of s mm; Assign random values to the grid of control points and smooth it with a Gaussian kernel; Resample the B-spline grid to obtain the DVF; Normalize the DVF linearly to be in the range $[-\theta, +\theta]$ along each axis.

mixed frequency: In this category, two different spatial frequencies are mixed together as follows: Create a single frequency DVF similar to the previous category; Create a random binary mask and multiply it with the single frequency DVF. Finally, smooth the DVF with a Gaussian kernel with a standard deviation of σ_B. σ_B is chosen relatively small to generate a higher spatial frequency in comparison with the smooth filled region; By varying the σ_B value and s in the filled DVF, different spatial frequencies will be mixed together.

respiratory motion: We simulate respiratory motion with three components similar to [35] as follows: Expansion of the chest in the transversal plane with a maximum scaling factor of 1.12; Transition of the diaphragm in cranio-caudal direction with a maximum deformation of θ; Random deformation using the single frequency method. In order to locate the diaphragm, an automatically detected lung mask is used.

identity: This category comprises only identity DVFs. Later, when creating the artificially deformed image, intensity augmentations will be added to the deformed image. Thus, the network will be capable of detecting no motion, while the intensity values might have changed slightly.

3.2.3.2 Artificial intensity models

We propose two intensity models to be applied on the fixed images:

Sponge intensity model: By assuming mass preservation over the lung deformation, a dry sponge model [73] is added to deformed image:

$$\boldsymbol{I_F}(\boldsymbol{x}) = \boldsymbol{I_F^{\text{clean}}}(\boldsymbol{x})[\boldsymbol{J_T}(\boldsymbol{x})^{-1}], \tag{3.1}$$

where $\boldsymbol{J}$ denotes the determinant of the Jacobian of the transformation.

Gaussian noise: A Gaussian noise with a standard deviation of $\sigma_N = 5$ is added to the deformed image in order to achieve more accurate simulation of real images.

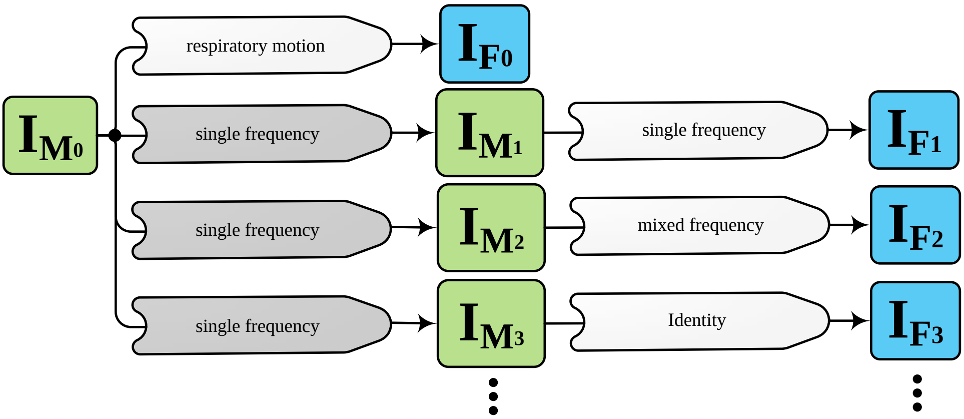
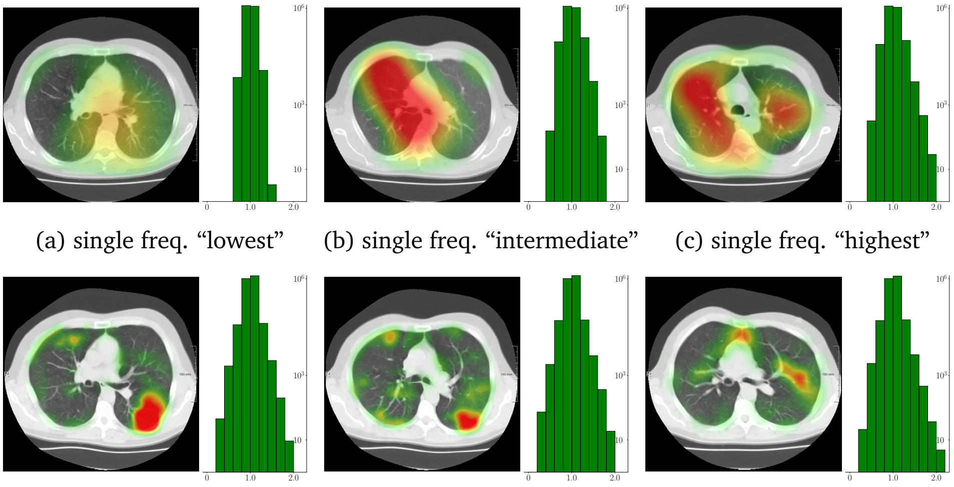


Figure 3.3: The generation of training pairs from a single input image I_{M0}. The input image is deformed slightly using the single frequency category, with the "lowest" settings (see Table 3.1), to generate moving images I_{Mi}. These are then each deformed and post-processed multiple times using all categories to generate fixed images I_{Fi}.

3.2.3.3 Extensive pair generation

For each single image in the training set, potentially a large number of artificial DVFs can be randomly generated. However, if this image is to be re-used for multiple DVFs, then for many training pairs we have the moving image unaltered. To tackle this problem, we also generate deformed versions of the original image (gray single frequency blocks in Fig. 3.3). A schematic design of utilizing artificial image pairs is depicted in Fig. 3.3. In this approach, the original image is only used once to generate the artificial image I_{F0}. Deformed versions of the original image I_{Mi} are used afterwards. Training pairs are thus $(I_{M0}, I_{F0}), (I_{M1}, I_{F1}), (I_{M2}, I_{F2}), \ldots$. The gray single frequency blocks in Fig. 3.3 have the same setting as single frequency "lowest" except that σ_N is set to 3 instead of 5. That is to avoid the accumulation of noise in the artificial images.

In total we generate 14 basis types of artificial DVFs: 5 single frequency, 4 mixed frequency, 4 respiratory motion and 1 identity. The precise settings of the parameters are available in Table 3.1 and examples are given in Fig. 3.4. The histograms of the Jacobians are also available in this figure. When the spatial frequency is increased, the Jacobian histograms will spread more, which shows that local relative volume changes are increased. The value of θ, the maximum artificial displacement along each axis, is chosen as 20, 15 and 7 for RegNet^4, RegNet^2 and RegNet^1, respectively.



(a) single freq. "lowest" (b) single freq. "intermediate" (c) single freq. "highest"

(d) mixed freq. "lowest" (e) mixed freq. "intermediate" (f) mixed freq. "high"

Figure 3.4: Examples of heat maps of generated artificial DVFs overlayed on the deformed images. We show three of the five spatial frequencies defined in Table 3.1. The histogram of the Jacobian determinant of each DVF is shown next to the sample image. As the spatial frequency increases, the histogram is more spread.

Table 3.1: DVFs with different spatial frequencies are obtained by varying the B-spline grid spacing s and the standard deviation of the Gaussian kernel σ_B. The maximum deformation along each axis θ only varies for each stage. When the spatial frequency is increased, the Jacobian histograms will spread more, which shows that local relative volume changes are increased (Fig. 3.4). S, M and R indicates single frequency, mixed frequency and respiratory motion.

Parameter	artificial DVF	lowest	low	intermediate	high	highest
θ (mm)	stage 1	3	7	7	7	7
	stage 2	5	15	15	15	15
	stage 4	7	20	20	20	20
s (mm)	S^1	[50, 50, 50]	[45, 45, 45]	[35, 35, 35]	[25, 25, 25]	[20, 20, 20]
	S^2	[60, 60, 60]	[50, 50, 50]	[45, 45, 45]	[40, 40, 40]	[35, 35, 35]
	S^4	[80, 80, 80]	[70, 70, 70]	[60, 60, 60]	[50, 50, 50]	[45, 45, 45]
	M^1	[50, 50, 50]	[40, 40, 40]	[25, 25, 35]	[20, 20, 30]	
	M^2	[60, 60, 60]	[50, 50, 40]	[40, 40, 80]	[35, 35, 80]	
	M^4	[80, 80, 80]	[60, 60, 60]	[50, 50, 50]	[45, 45, 60]	
	R^1	[50, 50, 50]	[45, 45, 45]	[35, 35, 35]	[25, 25, 25]	
	R^2	[60, 60, 60]	[50, 50, 50]	[45, 45, 45]	[40, 40, 40]	
	R^4	[80, 80, 80]	[70, 70, 70]	[60, 60, 60]	[50, 50, 50]	
σ_B	M^1	(5-10)	(5-10)	(5-10)	(5-10)	
	M^2	(7-12)	(7-12)	(7-12)	(7-12)	
	M^4	(10-15)	(10-15)	(10-15)	(10-15)	

3.2.4 Optimization

Optimization is done using the Adam optimizer with a learning rate of 0.001. The loss function consists of two parts. The first part is the Huber loss, which minimizes the difference between the ground truth $\boldsymbol{T}$ and the predicted DVF $\boldsymbol{T}'$ of the RegNet. The second part is a bending energy (BE) regularizer [6], which ensures smoothness of the displacement field:

$$\mathscr{C} = \text{Huber}\big(\boldsymbol{T}(\boldsymbol{x}), \boldsymbol{T}'(\boldsymbol{x})\big) + \gamma \cdot \text{BE}\big(\boldsymbol{T}'(\boldsymbol{x})\big), \tag{3.2}$$

where the Huber loss is defined as:

$$\text{Huber}(\boldsymbol{T}, \boldsymbol{T}') = \begin{cases} (\boldsymbol{T} - \boldsymbol{T}')^2, & |\boldsymbol{T} - \boldsymbol{T}'| \le 1, \\ |\boldsymbol{T} - \boldsymbol{T}'|, & |\boldsymbol{T} - \boldsymbol{T}'| > 1 \end{cases} \tag{3.3}$$

3.3 Experiments and results

3.3.1 Materials and ground truth

Three chest CT scan datasets are used in this study: The SPREAD [47], the DIR-Lab-4DCT [74] and the DIR-Lab-COPDgene dataset [75].

In the SPREAD database, 21 pairs of 3D chest CT images are available with a baseline and a follow-up image in each pair. The follow-up images are taken after 30 months. Both images are acquired in the inhale phase. Patients in this study are aged between 49 and 78 years old. The size of the images is approximately $446 \times 315 \times 129$ with a mean voxel size of $0.78 \times 0.78 \times 2.50$ mm. About 100 well-distributed corresponding landmarks were previously selected [73] semi-automatically on distinctive locations [48]. Two cases (12 and 19) are excluded because of the high uncertainty in the landmarks annotation [73].

In the DIR-Lab-COPDgene database, ten cases with severe breathing disorders are available in inhale and exhale phases. The average image size and the average voxel size are $512 \times 512 \times 120$ and $0.64 \times 0.64 \times 2.50$ mm, respectively. In each pair, 300 landmarks are annotated.

In the DIR-Lab-4DCT database ten cases are available. We use two phases of the available data: maximum inhalation and maximum exhalation. The size of the images is about $256 \times 256 \times 103$ with an average voxel size of $1.10 \times 1.10 \times 2.50$ mm.

Since the convolutional neural networks process the images in a voxel-based manner, all images are resampled to an isotropic voxel size of $1.0 \times 1.0 \times 1.0$ mm.

3.3.2 Evaluation measures

We use two measures to evaluate the performance of the proposed CNNs:

- **TRE:** The target registration error (TRE) defined as the mean Euclidean distance after registration between corresponding landmarks:

$$\mathrm{TRE} = \frac{1}{n}\sum_{i=1}^{n} \| \boldsymbol{T}'(\boldsymbol{x}_{Fi}) + \boldsymbol{x}_{Fi} - \boldsymbol{x}_{Mi} \|_2, \tag{3.4}$$

  where $\boldsymbol{x}_F$ and $\boldsymbol{x}_M$ are the landmark locations on the fixed and moving images, respectively.

- **Jac:** The Determinant of the Jacobian of the predicted DVF is calculated in order to measure relative changes in local volume. A very large (Jac $\gg 1$) or very small (Jac $\ll 1$) or negative Jac (Jac < 0) can indicate poor registration quality. We report the percentage of negative Jacobian as well as the standard deviation of the Jacobian inside the lung masks.

All of the assessments are performed on the real images.

3.3.3 Experimental setup

3.3.3.1 Training data

In the SPREAD database, 10 patients (20 images) are used for training, 1 patient (2 images) is in the validation set and 8 patients remain for the test set. From the DIR-Lab-COPD database, the first 9 cases (18 images) are used for training, and the remaining case (2 images) is used in the validation set. The entire DIR-Lab-4DCT database is used as an independent test set. The validation set is mainly used for tuning the hyper-parameters and selecting the best network design. In all evaluations, images are multiplied with the lung masks.

To generate training pairs, we use the 14 basis types of artificial generations (see Section 3.2.3.3). For each of the three networks, from each original image we generate 70 (5×basis), 42 (3×basis) and 28 (2×basis) artificial pairs in the first stage (RegNet^4), the second stage (RegNet^2) and the third stage (RegNet^1), respectively. Here we generate more images for more coarse stages, as these images are smaller.

In the training phase of the patch-based networks (MV, Uadv), the batch size is 15. The number of patches per pair is 5, 20, and 50 for stage 4, 2, and 1, respectively. The patch size is 101^3 and 105^3 for the U-Net-advanced and Multi-view design. When choosing samples, several balancing criteria are considered based on the magnitude of DVFs of the patches. An equal number of samples are selected from the range [0, 1.5), [1.5, 8) and [8, 20) mm for stage 4. For stage 2 and 1 these bins are selected as [0, 1.5), [1.5, 4), [4, 15) mm and [0, 2), [2, 7) mm, respectively. Training is run for 30 semi-epochs. All methods are trained with an additional data augmentation step, by adding Gaussian noise to all patches on the fly.

3.3.3.2 Software

In order to efficiently implement the artificial deformation and training phase, we utilize two processes. The task of the first process is to create artificial DVFs and deformed images and write them to disk. The second process has a multithreading paradigm which loads the data from disk and also handles the network training on the GPU.

The CNNs are implemented in Tensorflow [76]. Artificial DVFs are generated with the help of SimpleITK [51]. The code is publicly available via github.com/hsokooti/RegNet.

3.3.3.3 `elastix`

We compare the proposed CNN-based registration methods with conventional image registration, using `elastix` [52]. We used the following settings: metric: mutual information, optimizer: adaptive stochastic gradient descent, transform: B-spline, number of resolutions: 3, number of iterations per resolution: 500. For the public DIR-Lab-4DCT data, more conventional and CNN-based methods are compared with RegNet in Section 3.3.4.2.

3.3.4 Experiments

3.3.4.1 Architecture selection

In order to inspect the performance of the different architectures, an evaluation is performed on all pairs in the training and validation sets, i.e. half of the SPREAD data and the entire DIR-Lab-COPDgene data. We utilize the single and mixed category plus identity transform for artificial generations. Please note that the networks are trained with artificial image pairs i.e. during training both the fixed and moving images are deformed versions of the original images. For this evaluation however, we used the original non-deformed pairs, which the network has not seen.

As a first experiment, we train and validate the networks on the original image resolution only, i.e. without any multi-stage pipeline: see MV^1 and Uadv^1 in Section 3.2. It can be seen that the TREs of these networks are on the high end for both studies. Please note that due to high intensity variation in the baseline and follow-up images in the DIR-Lab-COPDgene database, the overall results are relatively poor. We discuss this issue later in Section 3.4.

In a second experiment, we train and test the networks on the lowest image resolution only, again without any multi-stage pipeline: see U^4, MV^4 and Uadv^4 in Table 3.2. Note that on the SPREAD data, the performance improved with respect to registration on the original resolution. The main reason is that the lowest resolution training set has the maximum deformation θ of 20 mm, whereas the maximum deformation was set to 7 mm in the original resolution training set (see Table 3.1). On the DIR-Lab-COPDgene data, similar results were obtained except for U^4.

Table 3.2: Quantitative results on the training and validation sets. The target registration error (TRE) is reported, together with the percenage of folding and the standard deviation of the Jacobian inside the lung masks. The networks are trained using artificial deformations from the single and mixed category plus identity (see Section 3.2.3.1). U, Uadv and MV represent the U-Net, U-Net-advanced and Multi-view design (see Section 3.2.2). A Wilcoxon signed-rank test is performed between U^4-Uadv^2-Uadv^1 and others, where † indicates a statistically significant difference with $p < 0.05$.

Network	SPREAD (case 1-11)			DIR-Lab-COPDgene		
	TRE (mm)	%folding	std(Jac)	TRE (mm)	%folding	std(Jac)
MV^1	3.86±4.32†	0.23±0.18	0.23±0.05	9.28±5.83†	0.24±0.07	0.29±0.03
Uadv^1	3.80±4.15†	0.24±0.20	0.28±0.06	9.65±6.19†	0.32±0.13	0.35±0.05
U^4	2.71±1.59†	0.00±0.00	0.09±0.02	10.2±6.00†	0.00±0.00	0.10±0.01
MV^4	2.30±1.80†	0.00±0.00	0.11±0.02	8.27±5.44†	0.00±0.00	0.14±0.02
Uadv^4	2.29±1.89†	0.00±0.00	0.08±0.01	8.60±5.50†	0.00±0.00	0.12±0.01
MV^4-MV^2	1.70±1.31†	0.00±0.01	0.18±0.03	6.67±5.53†	0.01±0.01	0.28±0.05
U^4-MV^2	1.71±1.23†	0.00±0.00	0.15±0.03	8.94±6.95†	0.03±0.02	0.27±0.06
Uadv^4-Uadv^2	1.69±1.34†	0.00±0.00	0.12±0.02	6.96±5.89†	0.00±0.00	0.21±0.04
U^4-Uadv^2	1.68±1.15†	0.00±0.00	0.11±0.02	8.54±6.91†	0.00±0.00	0.20±0.04
MV^4-MV^2-MV^1	1.63±1.30†	0.05±0.07	0.22±0.04	6.35±5.74†	0.44±0.24	0.38±0.08
U^4-MV^2-MV^1	1.60±1.20	0.02±0.03	0.19±0.04	8.65±7.27†	0.49±0.27	0.38±0.09
Uadv^4-Uadv^2-Uadv^1	1.60±1.23	0.03±0.04	0.22±0.03	6.45±6.39†	0.29±0.20	0.37±0.10
U^4-Uadv^2-Uadv^1	1.57±1.15	0.02±0.02	0.20±0.03	8.07±7.65	0.39±0.24	0.38±0.10

In the next experiment, we utilized two stages at image resolutions downsampled with a factor of 4 and then 2. In all four tested architectural combinations, the TRE results are better than the single stage networks in both studies, which shows that adding a second stage can improve the performance of RegNet.

Finally, when the original resolution is added to form a three-stage network a small improvement is observed in both studies. By comparing the final TRE results, it can be seen that the performance of all four network combinations are similar. For the remainder of the experiments, we choose the combination (U^4-Uadv^2-Uadv^1) as it obtained slightly better results on the SPREAD database. A Wilcoxon signed-rank test is performed between U^4-Uadv^2-Uadv^1 and other combination in Table 3.2. A statistically significant difference (with $p < 0.05$) between U^4-Uadv^2-Uadv^1 and all single stage and two stages combination can be observed.

3.3.4.2 Independent test set experiments

Now that we have selected the best network combination, we applied the U^4-Uadv^2-Uadv^1 pipeline on the independent test set (without retraining): 8 cases of the SPREAD

Table 3.3: Quantitative results of the SPREAD study in the training set (case 1 to case 11) and in the test set (case 13 to case 21). This experiment is performed with the network combination U^4-$Uadv^2$-$Uadv^1$. The target registration error (TRE) is reported, together with the percentage of folding and the standard deviation of the Jacobian inside the lung masks. S, M and R indicate single frequency, mixed frequency and respiratory motion, respectively (see Section 3.2.3.1). A Wilcoxon signed-rank test is performed between the B-spline registration and others. The symbol † indicates a significant difference between the average of TRE of B-spline registration and others, where † indicates a statistically significant difference with $p < 0.05$. The best method is shown in bold and the second best method is shown in green.

	`elastix`	`elastix` B-spline			RegNet				
	Affine				S	S+M	S+M+R	S	S
pair	TRE (mm)	TRE (mm)	%folding	std(Jac)	TRE (mm)	TRE (mm)	TRE (mm)	%folding	std(Jac)
case 1	8.77±2.76†	2.13±2.12	0.00	0.19	1.87±1.68	**1.78±1.66**	1.92±1.71	0.20	0.26
case 2	7.41±2.99†	1.48±1.18	0.00	0.11	1.44±0.92	**1.37±0.88**	1.54±1.03	0.03	0.20
case 3	4.34±1.89†	**1.66±1.13**	0.00	0.09	1.78±1.18	1.79±1.21	1.81±1.19†	0.00	0.15
case 4	11.4±3.44†	1.79±1.43	0.00	0.15	**1.70±1.41**	1.70±1.29	1.99±1.74	0.07	0.21
case 5	6.47±2.07†	**1.08±0.62**	0.00	0.09	1.15±0.76	1.17±0.70	1.24±0.78†	0.00	0.15
case 6	8.22±2.37†	2.06±1.37	0.00	0.14	1.98±1.44	**1.90±1.27**	1.92±1.22	0.02	0.21
case 7	5.51±1.38†	**1.50±1.11**	0.00	0.10	1.70±1.07†	1.63±1.22	1.51±1.10	0.00	0.17
case 8	3.67±2.31†	1.70±1.23	0.00	0.14	1.74±1.02	1.63±0.94	**1.53±0.84**	0.01	0.20
case 9	4.93±1.61†	**1.28±0.72**	0.00	0.09	1.35±0.72	1.41±0.87	1.51±0.77†	0.02	0.16
case 10	6.22±2.27†	**1.33±1.11**	0.00	0.10	1.40±0.90	1.40±0.87	1.43±0.85†	0.02	0.18
case 11	5.93±2.20†	**1.40±1.10**	0.00	0.13	1.49±1.15	1.44±1.19	1.51±1.08	0.03	0.21
Total	6.62±3.17†	1.58±1.29	0.00±0.00	0.12±0.03	1.60±1.17†	**1.57±1.15**	1.63±1.19†	0.04±0.05	0.19±0.03
case 13	12.5±15.8†	7.94±16.0	0.00	0.13	**7.37±14.0**	7.76±15.2	8.28±15.4	0.71	0.36
case 14	8.99±2.40†	1.86±1.19	0.00	0.11	**1.71±1.18**	2.08±1.81	2.25±1.76†	0.06	0.25
case 15	3.17±1.32†	**1.20±0.82**	0.00	0.11	1.39±0.86	1.29±0.82	1.33±0.84	0.00	0.18
case 16	8.94±1.84†	**1.30±0.80**	0.00	0.09	1.54±0.96	1.78±0.98†	1.96±1.01†	0.00	0.19
case 17	13.4±4.73†	**1.76±0.73**	0.00	0.09	2.89±3.66†	2.30±1.70†	3.37±3.43†	0.38	0.27
case 18	7.85±2.89†	1.65±1.41	0.00	0.15	**1.40±0.86**	1.60±1.16	1.71±1.04	0.02	0.21
case 20	4.43±2.14†	**1.31±0.90**	0.00	0.11	1.41±1.00	1.50±1.05†	1.52±0.97†	0.14	0.22
case 21	6.48±2.03†	**1.26±1.35**	0.00	0.09	1.36±1.19	1.33±1.36	1.36±1.47†	0.01	0.17
Total	8.16±6.76†	**2.21±5.86**	0.00±0.00	0.11±0.02	2.32±5.33†	2.39±5.64†	2.65±5.82†	0.19±0.24	0.24±0.06

database and the complete DIR-Lab-4DCT database. The results are given in Tables 3.3 and 3.4.

For the SPREAD database, the TRE results with affine and B-spline registration are compared with three versions of RegNet trained using the category “S” (single frequency plus identity), “S+M” (single frequency and mixed frequency plus identity) and “S+M+R” (single frequency plus mixed frequency and respiratory motion plus identity). Since there is no respiratory motion in the SPREAD data, adding respiratory motion did not improve the performance of the registration. Adding mixed frequencies did not change the results considerably: there was a small improvement for the cases 1-11, and slightly larger TREs for the cases 13 to 21. The percentage of folding inside the lung masks for the RegNet trained using “S” is also available in Table 3.3, which reports that the percentage of negative Jacobian are small in most cases, especially, when the TRE after affine registration is not very large. A Wilcoxon signed-rank test is performed between the `elastix` B-spline and other results. It can be seen that in

Table 3.4: Quantitative results on the DIR-Lab-4DCT study. This experiment is performed with the network combination of U^4-Uadv^2-Uadv^1. The target registration error (TRE) is reported, together with the percentage of folding and the standard deviation of the Jacobian inside the lung masks. S, M and R indicate single frequency, mixed frequency and respiratory motion, respectively (see Section 3.2.3.1). The result of [77] is the average of all respiratory phases per each case. The best method is shown in bold and the second best method is shown in green.

	elastix Affine	elastix B-spline	[78]	[77]	Vos et al. [23]	Eppenhof et al. [79]	Eppenhof et al. [79]-DIR	RegNet				
								S	S+M	S+M+R	S+M+R	S+M+R
pair	TRE (mm)	TRE (mm)	TRE (mm)	TRE (mm)	TRE (mm)	TRE (mm)	TRE (mm)	TRE (mm)	TRE (mm)	TRE (mm)	%folding	std(Jac)
case 01	3.02±2.13	1.21±0.71	**1.00±0.52**	1.20±0.60	1.27±1.16	1.45±1.06	-	1.09±0.51	1.12±0.54	1.13±0.51	0.00	0.10
case 02	3.76±3.20	1.39±1.27	**1.02±0.57**	1.19±0.63	1.20±1.12	1.46±0.76	1.24±0.61	1.08±0.89	1.06±0.57	1.08±0.55	0.00	0.12
case 03	5.92±3.64	2.44±2.11	**1.14±0.89**	1.67±0.90	1.48±1.26	1.57±1.10	-	1.23±0.69	1.23±0.75	1.33±0.73	0.00	0.14
case 04	9.01±4.53	2.16±2.16	**1.46±0.96**	2.53±2.01	2.09±1.93	1.95±1.32	1.70±1.00	1.47±0.95	1.62±1.09	1.57±0.99	0.00	0.18
case 05	3.95±2.85	3.02±3.22	1.61±1.48	2.06±1.56	1.95±2.10	2.07±1.59	-	**1.58±1.33**	1.60±1.33	1.62±1.30	0.00	0.14
case 06	10.7±6.80	3.33±3.30	**1.42±1.71**	2.90±1.70	5.16±7.09	3.04±2.73	-	4.56±7.06	4.95±6.91	2.75±2.91	0.03	0.25
case 07	11.1±7.43	6.16±6.33	**1.49±4.25**	3.60±2.99	3.05±3.01	3.41±2.75	-	6.10±7.10	5.00±6.35	2.34±2.32	0.03	0.24
case 08	12.0±6.59	9.36±9.30	**1.62±1.71**	5.29±5.52	6.48±5.37	2.80±2.46	-	6.54±8.51	6.18±7.01	3.29±4.32	0.01	0.22
case 09	7.89±3.83	3.31±2.74	**1.30±0.76**	2.38±1.46	2.10±1.66	2.18±1.24	1.61±0.82	2.02±2.25	1.84±1.93	1.86±1.47	0.00	0.17
case 10	6.87±6.12	2.72±3.43	**1.50±1.31**	2.13±1.88	2.09±2.24	1.83±1.36	-	2.82±4.93	2.44±3.85	1.63±1.29	0.00	0.19
Total	7.43±5.92	3.51±4.83	**1.36±1.01**	2.50±1.16	2.64±4.32	2.17±1.89	-	2.85±4.96	2.70±4.39	1.86±2.12	0.01±0.01	0.18±0.05

most cases there is no significant difference between B-spline registration and RegNet trained using “S” or trained using “S+M”.

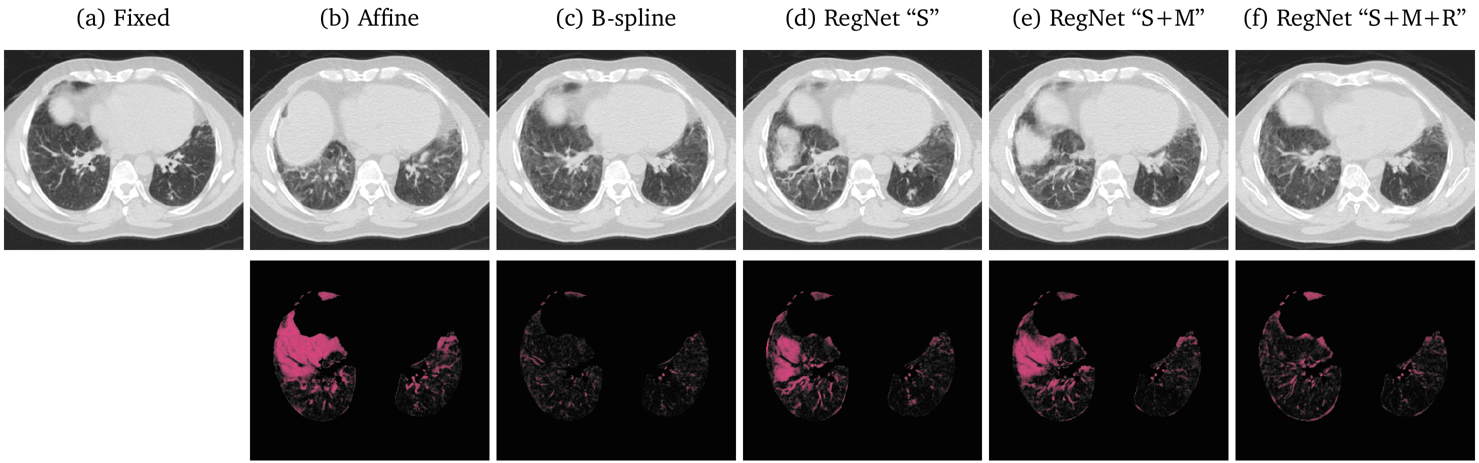
For the DIR-Lab-4DCT database, a comparison between RegNet and affine, B-spline (three resolutions), an advanced conventional registration method using sliding motion [78] and three other CNN-based methods [79, 23, 77] is available in Table 3.4. It can be seen that training with “S+M” improved performance slightly with respect to just “S”. Adding the respiratory motion category improved performance substantially, as these are inhale-exhale pairs; this is predominantly caused by the patients where the TRE after affine registration was still quite large. An example visualization is also available in Fig. 3.5, showing that adding the respiratory motion category can align images better in the diaphragm region. The advanced conventional registration method that leverages sliding motion [78] is still better than RegNet. Note that RegNet was not trained on the DIR-Lab-4DCT data, similar to [79, 77]. However, Vos et al. [23] and Eppenhof et al. [79] DIR methods were trained on the same database but using cross-validation to report the results. Also note that the results reported in [77] are averaged over all phases of DIR-Lab-4DCT (T00 to T10), while the results of other CNN methods (including RegNet) are reported between the maximum inhale and maximum exhale phase (T00 and T50). These reported results are therefore likely somewhat better than the results for T00 and T50 only.

3.3.4.3 Inference

At inference time, the patch size can be enlarged depending on the available GPU memory. For the U-Net-advanced design, the inference time of an image of size 101^3 and 269^3 voxels, is 0.02 s and 2.4 s, respectively, on our TITAN Xp (12 GB). An image of size 273^3 voxels took about 2.1 s to process for the Multi-view design. For the U-Net design we used the downsized image (by a factor of 4) of size 125^3 which took 0.02 s to be processed.

3.4 Discussion

In this paper we have shown that training a CNN with sufficiently realistic artificially generated displacement fields, can yield accurate registration results even in real cases. We utilized some randomly generated deformations (single and mixed frequencies) and a more realistic one (respiratory motion). We observed that even training with randomly generated deformations in the SPREAD study, the obtained TRE was on par with the B-spline registration (see Table 3.3). Adding more realistic DVFs (respiratory motion) in the DIR-Lab 4DCT study, improved the TRE results from 2.70±4.39 mm (“S+M”) to 1.86±2.12 mm (“S+M+R”) as can be seen in Table 3.4. In the case that sufficient realism was not added to the training, for instance in the DIR-Lab-COPDgene study in Table 3.2, the results were sub-optimal. Note that this dataset is challenging for conventional methods also. Anatomical structures in the baseline and follow-up



(a) Fixed (b) Affine (c) B-spline (d) RegNet “S” (e) RegNet “S+M” (f) RegNet “S+M+R”

Figure 3.5: Example results (top row) and difference images (bottom row) from DIR-Lab-4DCT study.

images of this database are quite different and the proposed intensity simulations in Section 3.2.3.2 did not cover this issue. A solution may be the addition of random intensity occlusions to the deformed images. Another interesting research direction is to learn realistic appearance and deformations from a database.

One of the major challenges of CNN-based image registration is capturing large DVFs, especially for patch-based methods. Using the whole image as an input might be very time consuming in the training phase. Based on our experiments, the maximum deformation that can be detected in patch with size $101 \times 101 \times 101$ by the U-Net-advanced is approximately 7 mm (the value of θ in Table 3.1). By enlarging the DVFs the Huber loss increased substantially. Please note that the maximum deformation θ is along each axis so the magnitude of a maximum deformation is $2 \times \sqrt{3} \times \theta$. Adding the original resolution makes the pipeline slower. However, based on the results in Table 3.2 it can be concluded that using two stages (U^4-$Uadv^2$, TRE: 1.68±1.15) can achieve similar results in comparison with three stages (U^4-$Uadv^2$-$Uadv^1$, TRE: 1.57±1.15). Similar to conventional registration, the best method on the first stage is not always the best in combination with others. In Table 3.2, the single U^4 is worse than $Uadv^4$ and MV^4. Conversely, the combination of U^4-$Uadv^2$-$Uadv^1$ obtains the best results. All in all, the differences between the architectures are relatively small and the performance of RegNet is more influenced by the artificial images used in training phase.

The performance of RegNet is very close to the conventional registration. However, B-spline registration is the best in the SPREAD test set (case 13 to 21; 2.21 ± 5.86 vs 2.32 ± 5.33) and the method of Berendsen et al. [78] performed better in the DIR-Lab-4DCT database (1.36 ± 1.01 vs 1.86 ± 2.12). On the contrary, the inference time of CNN approaches are much faster than the conventional methods. Potentially, by increasing the training data and generalizing the artificial generation like sliding motion, the performance of the RegNet can be improved.

In the current implementation of RegNet, all images are resampled to an isotropic voxel size $1.0 \times 1.0 \times 1.0$ mm. If resampling is not intended, it might be possible to simply multiply the output of RegNet by the voxel size. However, this approach is not very accurate because the spatial frequency of different voxel size might not be covered by the training data. A more accurate solution could be to include additional input of the voxel size to the network.

In principle, the proposed network design potentially can be utilized to predict the registration quality. Several methods are suggested by conventional learning using handcrafted features [55, 39] and a preliminary result by [44].

The proposed method can be trained and evaluated on other image modalities like brain MRI images. Potentially, the same network design and artificial generation excluding respiratory motion can be utilized.

The artificial generation can be enhanced if rib segmentation is available. Then, it is possible to incorporate rigid deformation outside of the rib and nonrigid deformations inside the rib. The network potentially can learn the relation between organs and rigidity of the deformations. More realistic and complex simulation like sliding motion of lungs [78] can also be added to the training images as it had a positive effect for non-learning based methods.

3.5 Conclusion

We proposed a 3D multi-stage CNN framework for chest CT registration. For training the network, we proposed models to generate artificial DVFs, and intensity models, to easily generate large quantities of paired images with a known spatial relation. We showed via multiple chest CT databases that this way of artificial training is very effective, with good results on real data. On the public DIR-Lab-4DCT database, we achieved the best results among the CNN approaches.

4

Quantitative Error Prediction of Medical Image Registration using Regression Forests

This chapter was adapted from:

H Sokooti, G Saygili, B Glocker, BP Lelieveldt, and M Staring. **Quantitative Error Prediction of Medical Image Registration using Regression Forests**, *Medical Image Analysis*, 2019.

Abstract

Predicting registration error can be useful for evaluation of registration procedures, which is important for the adoption of registration techniques in the clinic. In addition, quantitative error prediction can be helpful in improving the registration quality. The task of predicting registration error is demanding due to the lack of a ground truth in medical images. This paper proposes a new automatic method to predict the registration error in a quantitative manner, and is applied to chest CT scans. A random regression forest is utilized to predict the registration error locally. The forest is built with features related to the transformation model and features related to the dissimilarity after registration. The forest is trained and tested using manually annotated corresponding points between pairs of chest CT scans in two experiments: SPREAD (trained and tested on SPREAD) and inter-database (including three databases SPREAD, DIR-Lab-4DCT and DIR-Lab-COPDgene). The results show that the mean absolute errors of regression are 1.07 ± 1.86 and 1.76 ± 2.59 mm for the SPREAD and inter-database experiment, respectively. The overall accuracy of classification in three classes (correct, poor and wrong registration) is 90.7% and 75.4%, for SPREAD and inter-database respectively. The good performance of the proposed method enables important applications such as automatic quality control in large-scale image analysis.

4.1 Introduction

Image registration is the task of finding the optimal spatial transformation between two or more images. In most registration methods, no assessment of the registration quality is provided, and simply the result is returned. Evaluation of the registration is devolved to human experts, which is very time-consuming and prone to inter-observer errors as well as human fatigue [80]. Automatic quantitative error prediction of registration would decrease quality assessment time and can provide information about the registration uncertainty. Many medical pipelines are based on registered images and it is important to know the uncertainty of registration before continuing to a next phase in order to prevent accumulation of errors. For example, in online adaptive radiotherapy daily contouring of the tumor and organs-at-risk can be performed with the help of image registration [30]. In this task, quality assessment (QA) is mandatory to ensure patient safety. In addition, the accumulation of delivered dose over several treatment fractions is also impacted by the quality of registration [81, 82, 83]. Registration quality therefore has to be checked before the treatment starts. Visualizing the error of registration can also be directly helpful in medical applications before making a clinical decision. Smit et al. [31] localized autonomic pelvic nerves by registering a pre-operative MRI scan to an atlas model that includes nerve information. These nerves are not visible in the MRI scans and are prone to be damaged during a surgical procedure. Utilizing registration uncertainty yield better visualization of the autonomic nerves.

Refinement of registration is another important application of automatic error prediction. Muenzing et al. [32] improved registration by focusing only on regions with high registration error and discarding pixels which are aligned correctly. Registration refinement can also be done with the feedback of human experts by manually adding several corresponding landmarks [84].

Schlachter et al. [85] did a comprehensive study on visualization of registration quality with the help of three radiation oncologists on the DIR-Lab-COPDgene data, which has a slice thickness of 2.5 mm. The [average, maximum] TRE of the landmarks that were rated to be of acceptable registration quality with the conventional visualization method (checkerboard visualization and color blended) was [2.3, 6.9] mm, while with the best visualization method (histogram intersection) [1.8, 3.3] mm was achieved.

A few methods have been proposed to detect the misalignment of a pair of images with the purpose to refine the registration result. Rohde et al. [38] proposed to use the gradient of the cost function to detect which region in the image pair is poorly registered and potentially can be improved. Schnabel et al. [86] suggested to refine the registration by increasing the number of registration parameters in regions with high local entropy, or with high local variation in the intensity or with relatively steep cost function. In another work, analyzing the shape of the cost function around each voxel was used to estimate the confidence of registration [87]. Park et al. [37] used normalized local mutual information to find poorly aligned regions in order to increase the number of registration parameters. Forsberg et al. [88] utilized the outer product of the intensity gradient as an uncertainty measure in multi-channel diffeomorphic Demons registration. Although the mentioned metrics can be used to improve the image registration, it has not been shown how these metrics are correlated with the image registration error.

Several methods exploit continuous probabilistic image registration by utilizing Bayesian inference to achieve an intrinsic transformation uncertainty measure [89, 90]. However, it has been shown that there is no clear statistical correlation between transformation uncertainty and registration uncertainty [91]. The transformation and corresponding label (of a pair of images) are two random variables and it is not possible to quantify the uncertainty of the corresponding label by the summary statistics of the transformation. Another downside of these methods is that they can only be used for the specific paradigm of Bayesian registration.

Some methods are based on the consistency of multiple registrations between a group of images [92, 93], but these methods cannot be used in pairwise registrations.

In the stochastic approaches, Kybic [94] suggested to perform multiple registrations with random sampling of pixels with replacement. He found a correlation between the true registration error and the variation of the 2D translational parameters. The method was not extended to 3D and to nonrigid registration. Hub et al. [35] calculated the local mean square intensity difference multiple times by perturbing the B-spline grid. They showed that the maximum change of the dissimilarity metric in a local region is correlated with the registration error in that region. The drawback of this method is that it is not efficient in homogeneous areas [95]. In a related work they showed that the variance of the final deformation vector field (DVF) is related to the registration error [95], using the Demons algorithm. However, to find large misalignment a large search region is needed.

In this paper, we turn our attention to methods capable of *learning* the registration error allowing to take advantage of multiple features related to registration uncertainty within a single framework. Muenzing et al. [39] casted the registration assessment task to a classification problem with three categories (wrong, poor and correct registrations). In their method, they mostly utilize intensity-based features, except for the determinant of the Jacobian of the transformation. Although their training samples consist of manually selected landmarks, later they showed that assessing registration in all regions is possible by interpolation [32].

In our paper, instead of casting the uncertainty estimation task to a classification

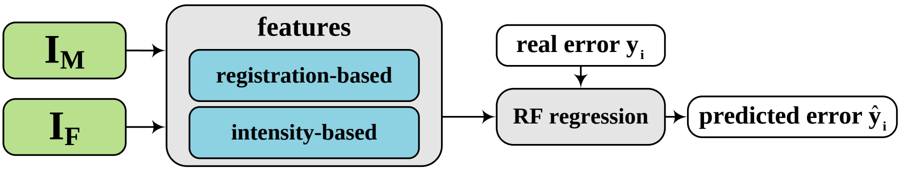


Figure 4.1: A block diagram of the proposed algorithm.

problem, we formulate it as a regression problem. To the best of our knowledge, in the field of continuous prediction of 3D registration error, Lotfi et al. [96] only tested their method on artificially deformed images. Recently Eppenhof et al. [44] estimated the registration error by utilizing convolutional neural networks. Only preliminary results were available for synthetic 3D data.

We explore several features related to the uncertainty of the registration transformation as well as related to intensity. All features are calculated in physical units, i.e. mm, which makes the system independent of voxel size. Finally, features are combined by using regression forests. The proposed method is applied and evaluated on chest CT scans. This work is an extension of [55] with updated methodology and substantially extended evaluation.

4.2 Methods

4.2.1 System overview

A block diagram of the proposed algorithm is shown in Fig. 4.1. The system has two inputs: a fixed image I_F and a moving image I_M. Several registration-based and intensity-based features are generated. A regression forests (RF) is then trained from all features to estimate the registration error.

The proposed system is trained to predict residual distances y (registration errors) obtained from a set of semi-automatically established corresponding landmarks. During evaluation, the prediction result $\hat{y}$ is compared with errors obtained from an independent set of ground truth landmarks, using cross-validation. The proposed system therefore estimates registration errors in physical units, i.e. mm. More information about the ground truth is available in Section 4.3.1. Details of the features are elaborated in Section 4.2.3.

4.2.2 Registration

Registration can be formulated as an optimization problem in which the cost function $\mathscr{C}$ is minimized with respect to $\boldsymbol{T}$:

$$\widehat{\boldsymbol{T}} = \arg\min_{\boldsymbol{T}} \mathscr{C}\big(\boldsymbol{T}; I_F, I_M\big), \tag{4.1}$$

where $\boldsymbol{T}$ denotes the transformation. The optimization is usually solved by an iterative method embedded in a multi-resolution setting. A registration can be initialized by an initial transform $\boldsymbol{T}^{\text{ini}}$.

4.2.3 Features and pooling

The features we used in our system, consist of several registration-based as well as intensity-based features. Some features are intrinsically capable to be calculated over differently sized local boxes, for others, a pool of features is created by computing local averages and maxima afterwards. The features used in this paper are listed in Table 4.1. We propose the following features:

4.2.3.1 Registration-based features

Variation of deformation vector field (std $\boldsymbol{T}$): The final solution of an iterative optimization problem can be influenced by the initial parameters. If in a region the cost function has multiple local minima or is semi-flat, a slight change in the initial parameters can lead to a different solution. In contrast, in areas where the cost function is well-defined, variations in the initial state are expected to have much less effect on the final solution. A flow chart of the described feature is available in Fig. 4.2. Given P random initial transformations $\boldsymbol{T}_i^{\text{ini}}$, $i \in \{1, \ldots, P\}$, that are used as initializations of the registration algorithm from Eq. (4.1), the variation in the final transformation results $\widehat{\boldsymbol{T}}_i$ is a surrogate for the precision of the registration. We propose to use the standard deviation std $\boldsymbol{T}$ of those final transformations as a feature:

$$\overline{\boldsymbol{T}} = \tfrac{1}{P} \sum \widehat{\boldsymbol{T}}_i, \tag{4.2}$$

$$\operatorname{std} \boldsymbol{T} = \sqrt{\tfrac{1}{P-1} \sum \| \widehat{\boldsymbol{T}}_i - \overline{\boldsymbol{T}} \|^2}. \tag{4.3}$$

In this work, the initial transformations $\boldsymbol{T}_i^{\text{ini}}$ are created by uniformly distributed offsets in the range $[-2, 2]$ mm to all B-spline coefficients. The offset range is chosen to be relatively small in comparison to the B-spline grid spacing in order to avoid unrealistic deformation. An example of std $\boldsymbol{T}$ in a synthetically deformed image is given in Fig. 4.3a.

Instead of perturbing the initial state of the registration, it is also possible to first perform the registration without any manipulated initial state, resulting in a transformation $\boldsymbol{T}^{\text{b}}$ [97]. Then, random offsets $\boldsymbol{T}_i^{\text{offset}}$ are added to $\boldsymbol{T}^{\text{b}}$ after which

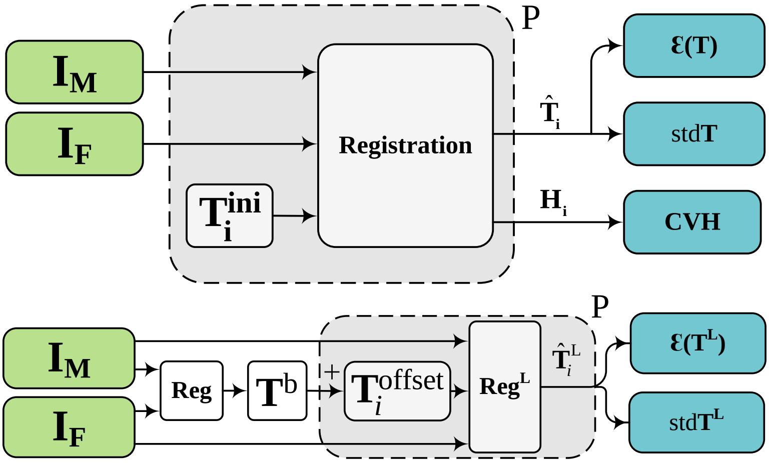
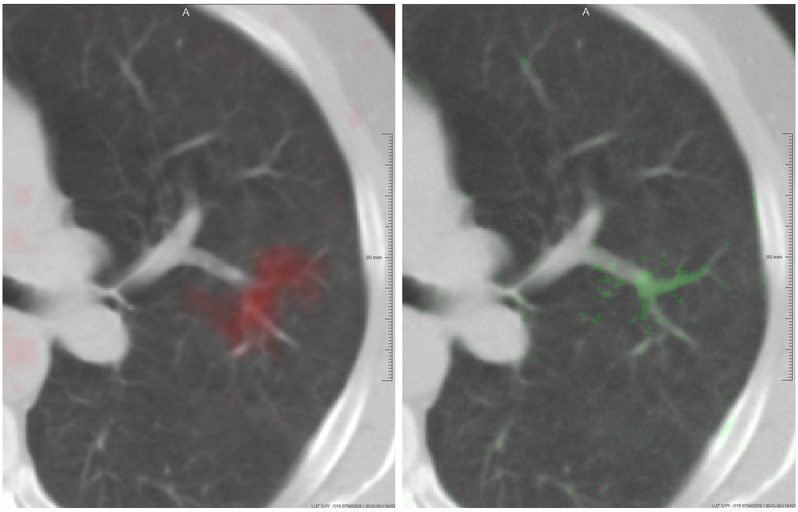


Figure 4.2: Multiple registrations are performed to create registration-based features. Either the initial transformation is varied, or the transformation after the base registration.



(a) std $\boldsymbol{T}$

(b) CVH

Figure 4.3: Visualization of std $\boldsymbol{T}$ and CVH in a synthetically deformed image. The deformed image is created by a random deformation vector field which is smoothed by a Gaussian kernel similar to [17].

another registration is performed, resulting in $\widehat{\boldsymbol{T}_i^{\mathrm{L}}}$. This is close to the work of Hub et al. [95], and approximately measures the concavity of the cost function. The feature

$\operatorname{std}\boldsymbol{T}^{\mathrm{L}}$ is then derived akin to Eq. (4.3):

$$\overline{\boldsymbol{T}^{\mathrm{L}}} = \tfrac{1}{P}\sum \widehat{\boldsymbol{T}_i^{\mathrm{L}}}, \tag{4.4}$$

$$\operatorname{std}\boldsymbol{T}^{\mathrm{L}} = \sqrt{\tfrac{1}{P-1}\sum \| \widehat{\boldsymbol{T}_i^{\mathrm{L}}} - \overline{\boldsymbol{T}^{\mathrm{L}}} \|^2}. \tag{4.5}$$

It is expected that $\operatorname{std}\boldsymbol{T}^{\mathrm{L}}$ is small in regions where the cost function is concave, as by adding small offsets $\boldsymbol{T}_i^{\mathrm{offset}}$ to the parameters, it can still move back to the previous optimal point. A flow chart of $\operatorname{std}\boldsymbol{T}^{\mathrm{L}}$ is shown in Fig 4.2. $\operatorname{std}\boldsymbol{T}^{\mathrm{L}}$ is calculated using the same setting as $\operatorname{std}\boldsymbol{T}$, except that only one resolution is used.

If the difference between $\overline{\boldsymbol{T}}$ and $\boldsymbol{T}^{\mathrm{b}}$ is relatively large, regions indicating a small $\operatorname{std}\boldsymbol{T}$ are still potentially regions of low registration quality. We then consider the bias $\mathscr{E}(\boldsymbol{T})$ and $\mathscr{E}(\boldsymbol{T}^{\mathrm{L}})$ as complementary features to $\operatorname{std}\boldsymbol{T}$ and $\operatorname{std}\boldsymbol{T}^{\mathrm{L}}$ computed by:

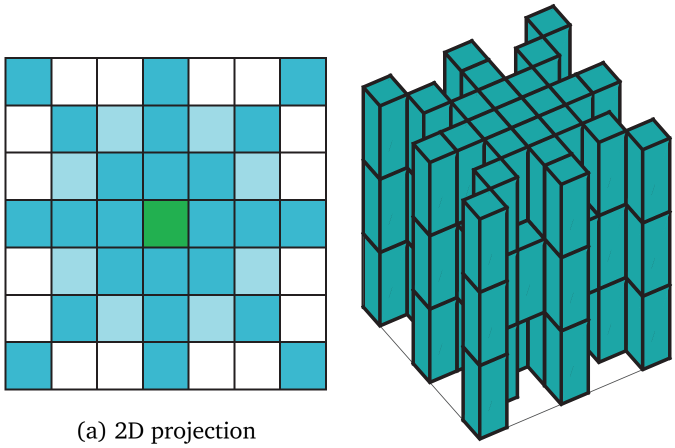
$$\begin{aligned} \mathscr{E}(\boldsymbol{T}) &= \| \boldsymbol{T}^{\mathrm{b}} - \overline{\boldsymbol{T}} \|, \\ \mathscr{E}(\boldsymbol{T}^{\mathrm{L}}) &= \| \boldsymbol{T}^{\mathrm{b}} - \overline{\boldsymbol{T}^{\mathrm{L}}} \|. \end{aligned} \tag{4.6}$$

Coefficient of variation of joint histograms (CVH): Multiple registration results can be used to extract additional information from the matched intensity patterns of the images. Given a fixed image I_F and a registration sub-result $I_M(\boldsymbol{T}_i)$, we calculate their joint histogram $\mathrm{H}_i, \forall i$. For identical sub-registrations, all resulting joint histograms are equal. Variation in the joint histograms implies registration uncertainty as a surrogate for registration error. The coefficient of variation of the joint histograms is calculated by dividing the standard deviation of all joint histograms over the average, $\overline{\mathrm{H}}$, of them. This normalization is done to compensate for large differences between the elements of $\overline{\mathrm{H}}$. We obtain the CVH in histogram space as follows:

$$\mathrm{CVH}^{\mathrm{B}\times\mathrm{B}} = \frac{\operatorname{std}\mathrm{H}}{\overline{\mathrm{H}} + \epsilon}, \tag{4.7}$$

where B is the number of histogram bins, and ϵ a constant to avoid division by zero. In the experiments we set ϵ to 5. The $\mathrm{CVH}^{\mathrm{B}\times\mathrm{B}}$ in histogram space is subsequently transferred to the spatial domain, by assigning voxels x with a particular intensity combination $\big(I_F(\boldsymbol{x}), I_M(\boldsymbol{T}^{\mathrm{b}}(\boldsymbol{x}))\big)$ the corresponding value from $\mathrm{CVH}^{\mathrm{B}\times\mathrm{B}}$, resulting in the final CVH feature with size equal to the fixed image. Note that the CVH can be used in a multi-modality setting, like the previous features. An example of the CVH on a synthetically deformed image is given in Fig. 4.3b.

Determinant of the Jacobian (Jac): Jac measures the relative local volume change. This can point to poor registration quality in case of very large ($\mathrm{Jac} \gg 1$) or very small ($\mathrm{Jac} \ll 1$) values, or discontinuous transformations in case of a negative value ($\mathrm{Jac} < 0$). In the experiments, the determinant of the Jacobian of $\boldsymbol{T}^{\mathrm{b}}$ is used.



(a) 2D projection

(b) 3D view

Figure 4.4: MIND search region. (a) The green cell indicates the center and darker blue cells indicate more accumulated cells in the projection view.

4.2.3.2 Intensity-based features

MIND: The Modality Independent Neighborhood Descriptor (MIND) was introduced by Heinrich et al. [14] in order to register multi-modal images. In this local self-similarity metric, a patch is considered to compare intensities between fixed and moving images. Finally, the sum of absolute differences between the MIND vector of I_F and that of $I_M(\boldsymbol{T}^{\mathrm{b}})$ is computed. We calculate MIND with a sparse patch including 82 voxels inside a $[7 \times 7 \times 3]$ box, which is approximately physically isotropic for the data used in the experiments (see Fig. 4.4).

Local normalized mutual information: Mutual information is used as an entropy-based similarity measure of two images. Similar to [39] we use the following definitions for local normalized mutual information:

$$\begin{aligned} \mathrm{NMI} &= \frac{H(I_F) + H(I_M(\boldsymbol{T}^{\mathrm{b}}))}{H\Big(I_F, (I_M(\boldsymbol{T}^{\mathrm{b}})\Big)}, \\ \mathrm{PMI} &= \frac{MI\Big(I_F, I_M(\boldsymbol{T}^{\mathrm{b}})\Big)}{min\Big\{H(I_F), H(I_M(\boldsymbol{T}^{\mathrm{b}}))\Big\}}. \end{aligned} \tag{4.8}$$

Both metrics are calculated over 8 differently sized boxes: [5, 10, 15, 20, 25, 30, 35, 40] mm. Two strategies for the selection of the number of bins are used, one uses a constant value $\mathrm{B_C}$, the other strategy depends on the number of samples $|\mathrm{B}| = log_2(n) + 1$, in which n is the number of samples in each box. The notations NMIS and PMIS indicate mutual information calculated with the latter strategy.

Table 4.1: An overview of the proposed features. Averages and maxima are taken over boxes of diameter [2, 5, 10, 15, 20, 25, 30, 35, 40] mm for the features: MIND, std $\boldsymbol{T}$, std $\boldsymbol{T}^{\mathrm{L}}$, CVH, $\mathscr{E}(\boldsymbol{T})$, $\mathscr{E}(\boldsymbol{T}^{\mathrm{L}})$ and Jac. Mutual information measures are calculated in boxes of [5, 10, 15, 20, 25, 30, 35, 40] mm. SID and GID are computed using Gaussian derivatives with standard deviations in the range [0.5, 1, 2, 4, 8, 16] mm.

Feature	N_f	
MIND	18	9 average boxes + 9 maxima boxes
MI	32	NMI, NMIS, PMI, PMIS calculated over 8 boxes
std $\boldsymbol{T}$	18	9 average boxes + 9 maxima boxes
std $\boldsymbol{T}^{\mathrm{L}}$	18	9 average boxes + 9 maxima boxes
CVH	18	9 average boxes + 9 maxima boxes
$\mathscr{E}(\boldsymbol{T})$	18	9 average boxes + 9 maxima boxes
$\mathscr{E}(\boldsymbol{T}^{\mathrm{L}})$	18	9 average boxes + 9 maxima boxes
Jac	18	9 average boxes + 9 maxima boxes
NC	8	calculated over 8 boxes
SID&GID	12	calculated over 6 sigma's

Modality-dependent features: In addition to the modality-independent features from above, we consider the use of several modality-dependent features. In the experiments we assess their contributed value. Similar to [39] the squared intensity difference (SID) and the gradient of intensity difference (GID) are computed using Gaussian (derivative) operators with standard deviations of [0.5, 1, 2, 4, 8, 16] mm. Normalized correlation (NC) is calculated within boxes of size [5, 10, 15, 20, 25, 30, 35, 40] mm akin to [39].

4.2.3.3 Pooling

In order to reduce discontinuities and improve interaction with other features, the total set of features is increased by generating a pool from those mother features by calculating averages and maxima over them using differently sized boxes. The features MI, SID, GID and NC are inherently computed over differently sized local regions. The features MIND, std $\boldsymbol{T}$, std $\boldsymbol{T}^{\mathrm{L}}$, CVH, $\mathscr{E}(\boldsymbol{T})$, $\mathscr{E}(\boldsymbol{T}^{\mathrm{L}})$ and Jac are calculated in a voxel-based fashion, and then pooled afterwards. Average and maximum pooling is performed with box sizes of $[2, 5, 10, 15, 20, 25, 30, 35, 40]$ mm. As a result, for each feature we obtain a pool of 18 features: 9 from box averages and 9 from box maxima. The average-pooling is done efficiently by the help of integral images introduced by Viola et al. [98]. A list of the proposed mother features together with the number of derived features N_f are given in Table 4.1.

4.2.4 Regression forests

Random forests were introduced by Breiman [99] by extending the idea of bagging. The forests consist of several weak learners (trees) which are combined in an efficient fashion. Each tree is started from a node and continues splitting until reaching certain criteria. In contrast to bagging, splitting is performed with a random subset of features which makes the training phase faster and reduces correlation between trees, consequently decreasing the forest error rate. The reason that we chose the random forest is that it can handle data without preprocessing. For instance rescaling of data, outlier removal and selection of features are not necessary in random forests. In addition, random forest are efficient to train and fast at runtime.

Random forests have the capability to calculate the importance of each feature with a little additional computation, which shows the contribution of each feature to the forest. Training of each tree is based on a bootstrap of all samples, and the so-called out-of-bootstrap samples Ω are used to compute the importance of a feature x_i. Importance is then defined as the difference between the mean square error (MSE) before and after a permutation of this feature:

$$\mathrm{Imp}(x_i) = \frac{1}{N_t} \sum_{t=1}^{N_t} \left(\underset{j \in \Omega}{\mathrm{MSE}} \left(\hat{y}_{\pi_i j}, y_j \right) - \underset{j \in \Omega}{\mathrm{MSE}} \left(\hat{y}_j, y_j \right) \right), \tag{4.9}$$

where y_j is the real value, $\hat{y}_j$ the predicted value from the regression, $\hat{y}_{\pi_i j}$ the predicted value when permuting feature i, and N_t the number of trees.

In this work, random forests are trained with different combinations of the proposed features (see Table 4.1). The dependent variable y is the registration error in mm, which is described in Section 4.3.1.

4.3 Experiments and results

4.3.1 Materials and ground truth

The SPREAD [47] DIR-Lab-4DCT [74] and DIR-Lab-COPDgene [75] databases have been used in this study. In the SPREAD study, there are 21 pairs of 3D follow-up lung CT images. Each patient in this database has a baseline and a follow-up image (which is taken after 30 months) both in inhale phase. The age of the patients ranges from 49 to 78 years old. The average size of the images is $446 \times 315 \times 129$ with an average voxel size of $0.78 \times 0.78 \times 2.50$ mm. In each pair of images, about 100 well-distributed corresponding landmarks were previously selected [73] semi-automatically on distinctive locations [48].

From the DIR-Lab-4DCT data, five cases (4DCT1 to 4DCT5) are selected with each five phases between maximum inhalation and exhalation. The average image size is $256 \times 256 \times 103$ with an average voxel size of $1.10 \times 1.10 \times 2.50$ mm. Each scan has 75 corresponding landmarks annotated. Ten cases with severe breathing disorders are available via the DIR-Lab-COPDgene database. The images are taken in inhale and exhale phases. In total, 300 landmarks are annotated. The average image size and the average voxel size are $512 \times 512 \times 120$ and $0.64 \times 0.64 \times 2.50$ mm, respectively.

Accuracy of the registration can be defined as the residual Euclidean distance after registration between the corresponding landmarks:

$$y = \|\boldsymbol{T}^{\mathrm{b}}(\boldsymbol{x}_F) - \boldsymbol{x}_M\|_2, \tag{4.10}$$

with $\boldsymbol{x}_F$ and $\boldsymbol{x}_M$ the corresponding landmark locations. Based on the idea that the registration error is smooth, we include voxels from a small local neighborhood around the landmarks to increase the total set of available landmarks. In this small neighborhood we assume that the registration error is equal to the error at the center of the neighborhood. This assumption seems reasonable for smooth transformations and within a small region. The neighborhood size is chosen as $10 \times 10 \times 7.5$ mm, which is approximately equivalent to the final grid spacing of the B-spline registration (see Fig. 4.5).

The core software is written in Python. The feature pooling is performed with a C++ program [100] and the regression forest is calculated with the help of the Scikit-learn package [101]. All registrations are performed by `elastix` [52]. Detailed registration setting can be found in the `elastix` parameter file database (elastix.isi.uu.nl, entry par0049). The code is publicly available via github.com/hsokooti/regun.

4.3.2 Evaluation measures

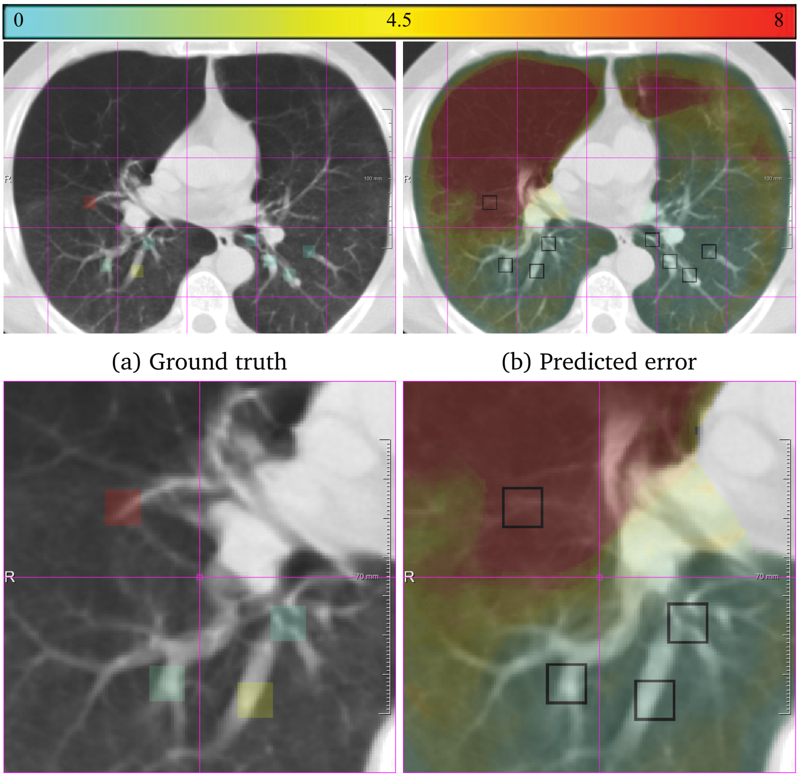
In the SPREAD database, we employ 10 cross-validations by randomly splitting the data in 15 image pairs for training and the remaining 6 pairs for testing. To evaluate the regression performance, the mean absolute error (MAE) of the real registration error y_i and the estimated one $\hat{y}_i$ is calculated over the neighborhood of the landmarks by:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} |\hat{y}_i - y_i|. \tag{4.11}$$

To further detail the regression performance, the MAE is subdivided into three categories: $\mathrm{MAE_c}$, $\mathrm{MAE_p}$ and $\mathrm{MAE_w}$ with y in $[0,3)$, $[3,6)$ and $[6,\infty)$ mm, corresponding to correct, poor and wrong registration, similar to Muenzing et al. [39]. We then do the same for $\hat{y}_i$, and report the accuracy and F1 score for classifying the registration error in these three categories.

4.3.3 Parameter selection

The RF is trained using 100 trees with a maximum tree depth of 9, while at least 5 samples remain in the leaf nodes. At each splitting node, m features are randomly selected. We set m to the square root of the total number of features in that experiment, which performed slightly better than $m =$ (number of features)$/3$ [102]. The total number of registrations P is chosen as 20 to ensure that the estimation of $\mathrm{std}\,\boldsymbol{T}$ does not change considerably when increasing the number of registrations [55].



(a) Ground truth

(b) Predicted error

(c) Magnification of (a)

(d) Magnification of (b)

Figure 4.5: Example data from the SPREAD dataset. The left column (a,c) shows the fixed image with the ground truth registration error overlaid in color. The square boxes around each landmark are given the same error as the error at the landmark. The right column (b,d) shows the moving image after registration with the registration error predicted by the proposed method overlaid in color. (c) and (d) are zoomed in versions of (a) and (b).

4.3.4 Reference registration error set

For the SPREAD and the DIR-Lab-4DCT study, registrations are based on free-form deformations by B-splines [6]. The cost function is mutual information, which is optimized by adaptive stochastic gradient descent. We used three resolutions with a final B-spline grid spacing of $[10, 10, 10]$ mm. We collect samples by performing four different registrations using 20, 100, 500 and 2000 iterations, respectively. All other registration settings remain the same in these registrations. By varying the number of iterations we increase the variation in the samples, as well as the training size. Table 4.2 gives the distribution of reference registration errors in each database. As expected, increasing the number of iterations shifts the distribution towards the

Table 4.2: Distribution of the reference registration errors in each database, used during testing.

Database-iters	correct		poor		wrong		total
SPREAD 20	848789	(84.1%)	102837	(10.2%)	58059	(5.8%)	1009685
SPREAD 100	904796	(89.6%)	66467	(6.6%)	38422	(3.8%)	1009685
SPREAD 500	925840	(91.7%)	51910	(5.1%)	31935	(3.2%)	1009685
SPREAD 2000	935676	(92.7%)	46170	(4.6%)	27839	(2.8%)	1009685
SPREAD together	3615101	(89.5%)	267384	(6.6%)	156255	(3.9%)	4038740
DIR-Lab-4DCT 20	521481	(84.5%)	71282	(11.5%)	24543	(4.0%)	617306
DIR-Lab-4DCT 100	540989	(87.6%)	61131	(9.9%)	15186	(2.5%)	617306
DIR-Lab-4DCT 500	553757	(89.7%)	53067	(8.6%)	10482	(1.7%)	617306
DIR-Lab-4DCT 2000	561909	(91.0%)	46679	(7.6%)	8718	(1.4%)	617306
DIR-Lab-4DCT together	2178136	(88.2%)	232159	(9.4%)	58929	(2.4%)	2469224
DIR-Lab-COPD ANTsBSplineSyN	2643	(88.1%)	184	(6.1%)	173	(5.8%)	3000
DIR-Lab-COPD `elastix-advanced`	2420	(80.7%)	259	(8.6%)	321	(10.7%)	3000

Table 4.3: Distribution of the reference registration errors, used during training.

Database	correct		poor		wrong		total
SPREAD together	589854	(58.0%)	270523	(26.6%)	156881	(15.4%)	1017258
DIR-Lab-4DCT together	328055	(53.0%)	232499	(37.5%)	58929	(9.5%)	619483

“correct" registration category. The maximum registration error is 81.8 mm in the SPREAD database, 17.6 mm in the DIR-Lab-4DCT database.

Since the a priori distribution of registration errors is imbalanced, with much more samples in the “correct" category, we perform the following balancing step during training. For landmarks that fall in the category “correct", we only add samples from a smaller neighborhood of $5 \times 5 \times 2.5$ mm instead of the $10 \times 10 \times 7.5$ mm neighborhoods used for landmarks in the categories “poor" and “wrong". The distribution of reference registration errors of the training samples is shown in Table 4.3.

For the DIR-Lab-COPDgene study, more advanced settings of the registration are used. In this experiment, samples are taken only on the landmark locations. More details are given in Section 4.3.5.8. The maximum registration error in this data is 31.5 mm.

4.3.5 Experiments

4.3.5.1 Single feature performance in SPREAD

The proposed features are described in Section 4.2.3 and summarized in Table 4.1. To investigate the strength of the individual features, we trained the random forest with only a single feature with pooling. By comparing the MAE results in Table 4.4, it can be seen that MIND, std $\boldsymbol{T}^{\mathrm{L}}$ and SID&GID are the best single features in the categories

Table 4.4: Regression results for single features on the SPREAD database. The columns indicate the number of features (N_f), the mean absolute error (MAE), the accuracy (Acc) and the F1 score. The sub-indices c, p and w correspond to correct [0,3), poor [3,6) and wrong [6, ∞) mm classes, respectively.

	N_f	MAE	MAE_c	MAE_p	MAE_w	Acc	F1_c	F1_p	F1_w
MIND	18	1.10 ± 1.97	0.76 ± 0.72	1.59 ± 1.39	6.50 ± 5.88	89.8	94.9	34.1	83.0
MI	32	1.20 ± 1.88	0.89 ± 0.71	1.53 ± 1.14	6.30 ± 5.58	87.9	93.9	30.1	79.9
std $\boldsymbol{T}$	18	1.59 ± 2.79	1.15 ± 1.78	2.98 ± 4.06	7.60 ± 6.12	85.5	92.7	22.4	64.4
std $\boldsymbol{T}^\text{L}$	18	1.51 ± 2.40	1.11 ± 1.34	2.49 ± 3.05	7.32 ± 5.79	86.7	93.4	18.3	70.7
CVH	18	1.93 ± 3.29	1.49 ± 2.22	1.82 ± 2.00	9.80 ± 7.19	75.2	87.2	16.9	37.0
$\mathscr{E}(\boldsymbol{T})$	18	2.00 ± 2.80	1.61 ± 1.76	2.18 ± 3.12	8.52 ± 6.48	69.8	82.8	17.0	43.5
$\mathscr{E}(\boldsymbol{T}^\text{L})$	18	1.68 ± 2.85	1.19 ± 1.71	3.19 ± 3.28	8.34 ± 6.74	84.4	92.6	11.7	54.8
Jac	18	2.15 ± 3.15	1.72 ± 1.90	1.91 ± 2.27	10.03 ± 6.97	68.2	83.7	13.0	31.4
NC	8	1.38 ± 2.89	0.90 ± 0.71	1.70 ± 1.68	9.41 ± 9.15	88.2	94.3	28.5	77.0
SID&GID	12	1.30 ± 2.12	0.94 ± 0.90	1.82 ± 1.63	6.95 ± 6.02	89.9	95.1	24.9	74.3

Intensity, Registration and Modality-dependent, respectively.

4.3.5.2 Combined features performance

Instead of using only a single feature, several combinations of features are used to build the RFs:

- **Intensity:** Combination of all modality-independent intensity features: MIND and MI (50 features).
- **Registration:** Combination of all registration features: std $\boldsymbol{T}$, std $\boldsymbol{T}^\text{L}$, CVH, $\mathscr{E}(\boldsymbol{T})$, $\mathscr{E}(\boldsymbol{T}^\text{L})$ and Jac (108 features).
- **Combined:** Combination of both intensity and registration features (158 features).

All results are available in Table 4.5. By combining features from both the registration and modality-independent intensity category, improvements were obtained in all evaluation measures.

The result of the regression with combined features is detailed in Fig. 4.6(a), which shows the real error (solid blue line) against the predicted error, sorted from small to large. In Fig. 4.6(b) we grouped the real errors in the three categories, each category showing a box-plot of the predicted errors. Intuitively, a smaller overlap between the boxes represents a better regression.

4.3.5.3 Including modality-dependent features

We consider adding the combination of three modality-dependent features to the combined feature set (Combined+MD): NC, SID and GID. In both databases, if we add the modality-dependent features (see Table 4.5), negligible differences are observed. Therefore, to keep the feature set small and modality-independent, we select the

Table 4.5: Regression results for groups of features on the SPREAD database. The columns indicate the number of features (N_f), the mean absolute error (MAE), the accuracy (Acc) and the F1 score. The sub-indices c, p and w correspond to correct [0,3), poor [3,6) and wrong [6, ∞) mm classes, respectively. MD, NN and LR stands for modality dependent, neural networks and linear regression, respectively.

	N_f	MAE	MAE$_\mathrm{c}$	MAE$_\mathrm{p}$	MAE$_\mathrm{w}$	Acc	F1$_\mathrm{c}$	F1$_\mathrm{p}$	F1$_\mathrm{w}$
Intensity	50	1.09 ± 1.88	0.77 ± 0.68	1.49 ± 1.26	6.20 ± 5.68	90.3	95.1	35.7	83.6
Registration	108	1.32 ± 2.35	0.90 ± 1.04	2.10 ± 2.71	7.76 ± 6.01	90.0	95.1	31.5	78.4
Combined	158	1.07 ± 1.86	0.76 ± 0.65	1.47 ± 1.22	6.12 ± 5.64	90.7	95.4	38.1	84.4
Combined-no pooling	8	1.24 ± 2.22	0.85 ± 0.73	1.72 ± 1.64	7.39 ± 6.62	89.4	94.8	32.6	79.1
Combined+MD	178	1.07 ± 1.83	0.76 ± 0.65	1.46 ± 1.20	5.95 ± 5.59	90.7	95.4	38.3	84.5
Combined (LR)	158	1.86 ± 2.03	1.58 ± 1.34	2.47 ± 2.21	6.12 ± 4.97	77.3	87.3	17.0	67.6
Combined (NN)	158	1.13 ± 2.07	0.74 ± 0.70	1.81 ± 1.67	7.08 ± 5.88	89.8	95.0	31.2	79.6

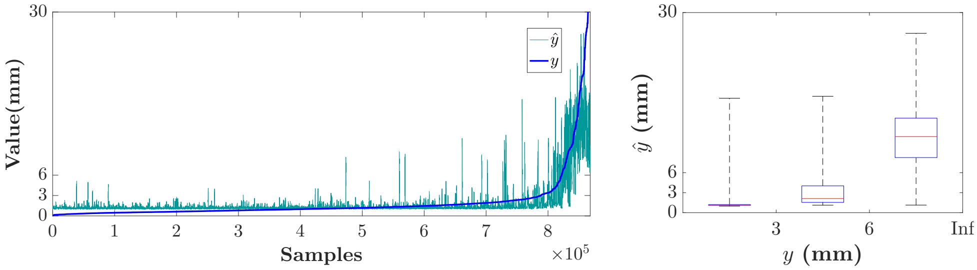


Figure 4.6: Real (y) vs predicted registration error ($\hat{y}$) for Combined features in the SPREAD database. (a) The real error (solid blue line y) against the predicted error ($\hat{y}$), sorted from small to large. In (b) we grouped the real errors in the three categories, each category showing a box-plot of the predicted errors.

“combined features" class without the modality-dependent features as the final system in the remainder of this paper.

4.3.5.4 The effect of pooling

To examine the effect of pooling, we perform an experiment without pooling on the combined feature set. We only calculate PMIS within a box size of 15 mm in this experiment. From Table 4.5 the benefit of pooling can be observed.

4.3.5.5 Alternative regression methods

In this section, we compare RF regression with linear regression (LR) and neural networks (NN). Feature normalization is done for both regressors. We utilized neural networks with three hidden layers of 1024, 512 and 256 units each. ReLU is used as an activation function and Huber is utilized as a loss function. Table 4.5 gives the results of these experiments. The performance of neural networks is on par with random forests. However, the results of linear regression are not comparable to that of random forests, both in MAE and accuracy.

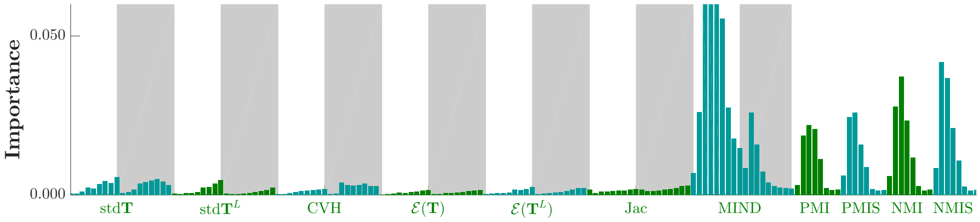


Figure 4.7: Feature importance of the SPREAD combined experiment. White areas correspond to box averages, while shaded areas correspond to box maxima.

Table 4.6: Leave one feature out results of SPREAD data. The columns indicate the number of features (N_f), the mean absolute error (MAE), the accuracy (Acc) and the F1 score. The sub-indices c, p and w correspond to correct [0,3), poor [3,6) and wrong [6, ∞) mm classes, respectively.

	N_f	MAE	MAE_c	MAE_p	MAE_w	Acc	F1_c	F1_p	F1_w
Combined	158	1.07 ± 1.86	0.76 ± 0.65	1.47 ± 1.22	6.12 ± 5.64	90.7	95.4	38.1	84.4
−MIND	140	1.18 ± 1.96	0.83 ± 0.66	1.56 ± 1.50	6.70 ± 5.69	90.2	95.1	36.2	83.0
−MI	126	1.10 ± 1.98	0.75 ± 0.67	1.54 ± 1.30	6.66 ± 5.84	90.6	95.3	37.0	84.2
− std $\boldsymbol{T}$	140	1.08 ± 1.86	0.76 ± 0.65	1.46 ± 1.18	6.14 ± 5.65	90.7	95.3	38.1	84.3
− std $\boldsymbol{T}^\text{L}$	140	1.08 ± 1.89	0.76 ± 0.65	1.46 ± 1.22	6.21 ± 5.73	90.6	95.3	38.3	83.7
−CVH	140	1.07 ± 1.81	0.75 ± 0.65	1.46 ± 1.21	6.06 ± 5.98	90.7	95.4	38.4	84.3
$-\mathscr{E}(\boldsymbol{T})$	140	1.07 ± 1.86	0.76 ± 0.65	1.46 ± 1.21	6.13 ± 5.64	90.7	95.4	38.2	84.5
$-\mathscr{E}(\boldsymbol{T}^\text{L})$	140	1.08 ± 1.85	0.76 ± 0.65	1.47 ± 1.22	6.12 ± 5.61	90.6	95.3	37.5	84.3
−Jac	140	1.08 ± 1.87	0.76 ± 0.65	1.49 ± 1.31	6.06 ± 5.72	90.7	95.4	37.9	84.8

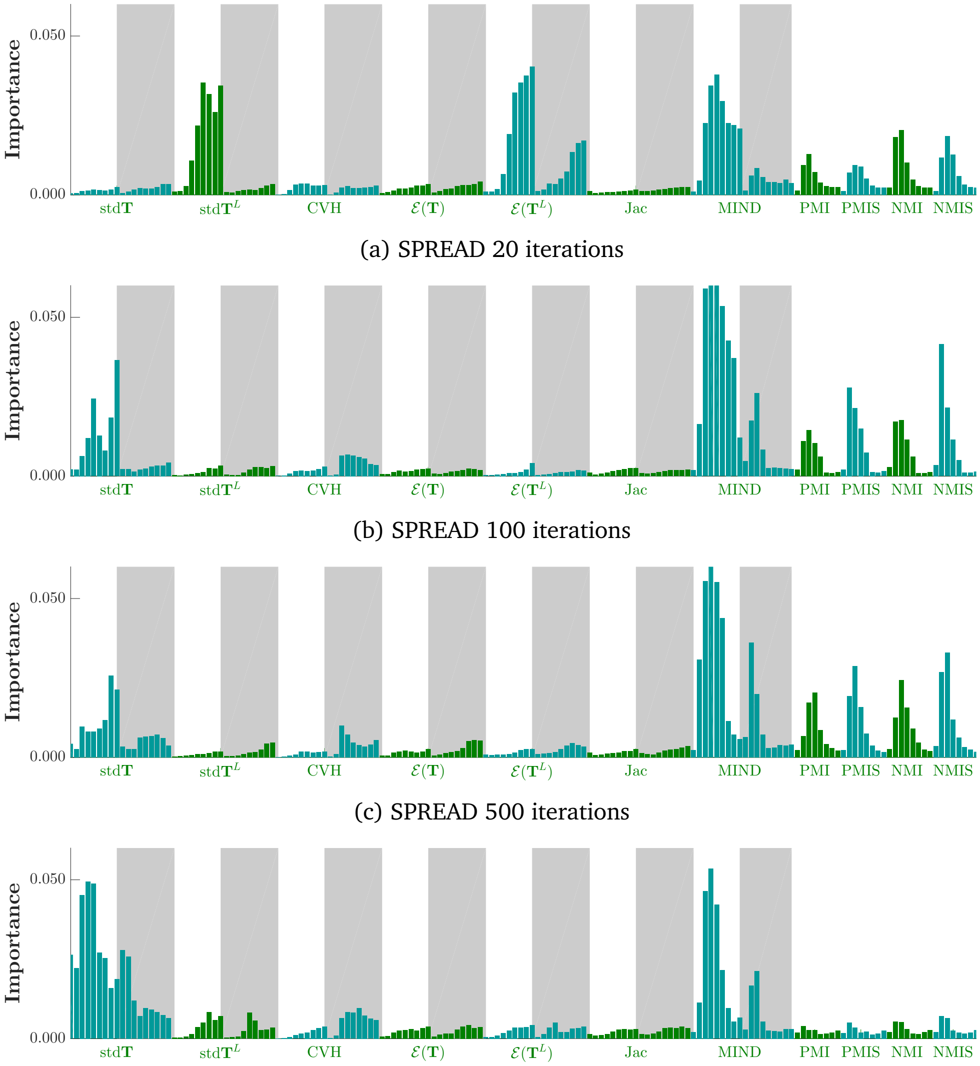
4.3.5.6 Feature importance

The feature importance, see Eq. (4.9), is displayed in Fig. 4.7. It shows that MIND and MI are the features contributing most to the RF performance, followed by std $\boldsymbol{T}$, std $\boldsymbol{T}^\text{L}$ and CVH.

The feature importance using a different number of iterations is shown in Fig. 4.8. The contribution of all intensity features stay the same in all experiments, while some of the registration features contribute differently with respect to the number of iterations. For instance, the importance of std $\boldsymbol{T}$ and CVH increase with increasing the number of iterations. The features std $\boldsymbol{T}^\text{L}$ and $\mathscr{E}(\boldsymbol{T}^\text{L})$ play important roles when the number of iterations is not enough for registration convergence.

4.3.5.7 Excluding a single feature

To further investigate the importance of the several features, we additionally perform an experiment where we leave one feature out of the combined feature set. The results are reported in Table 4.6. In these experiments, feature redundancy can be found. For instance, MI has a large importance values in random forests, but if we leave that feature out, other features can compensate for that.



(a) SPREAD 20 iterations

(b) SPREAD 100 iterations

(c) SPREAD 500 iterations

(d) SPREAD 2000 iterations

Figure 4.8: Feature importance of the SPREAD combined experiment with different iterations. The contribution of all intensity features stay the same in all experiments, while some of the registration features contribute differently with respect to the number of iterations. White areas correspond to box averages, while shaded areas correspond to box maxima.

4.3.5.8 Inter-database validation

To study the generalizability of the proposed system, instead of cross-validation on a single database, we perform training on the DIR-Lab-4DCT database and test it on the

SPREAD database. As mentioned before, the SPREAD database consists of only inhale images but the DIR-Lab-4DCT database has images from inhale to exhale phases. Therefore, this makes the DIR-Lab-4DCT more suitable for training. The result of this experiment is available in Table 4.7. Once more, we can draw the conclusion that by combining both intensity and registration-based features, the regression performance can be improved. In contrast to the SPREAD experiment, this time it is observed that the registration features perform better than the intensity features.

To further evaluate the generalizability of the proposed method, we test it for different registration methods on a third independent test set, the DIR-Lab-COPDgene dataset. The regression forest is trained on a combination of the SPREAD and DIR-Lab-4DCT data. We evaluate two registration algorithms that achieved excellent performance in the EMPIRE10 challenge [80], i.e. the ANTs registration package [103, 104] and `elastix` with advanced settings [105].

Prior to deformable registration we perform an affine registration using 5 resolutions and utilizing torso masks. For the deformable registration we use settings similar to the ones used in the EMPIRE10 challenge, specifically:

ANTs-BSplineSyN: With respect to the EMPIRE10 challenge we increased the number of iterations to 1000 for each of the 4 resolutions, using a 10% sampling rate. This improved the performance on our data and considerably reduced the calculation time. As suggested in [104], several preprocessing steps are used, including masking out the lungs, and inverting the image intensities and rescaling them between 0 and 1. Further settings include: registration model: symmetric diffeomorphic; dissimilarity metric: local cross correlation; number of resolutions: 4; maximum number of iterations: 1000; sampling: 10% random samples; convergence threshold: 1e-6. The average TRE on DIR-Lab-COPDgene is 1.90 ± 2.86 mm.

`elastix-advanced`: Settings are adopted from [105]. The most important ones are: registration model: B-spline; dissimilarity metric: normalized correlation; number of resolutions: 6; number of iterations: 1000; sampling: 2000 random samples; B-spline grid spacing: [5, 5, 5] mm. The average TRE with this setting is 3.39 ± 4.30 mm on the DIR-Lab-COPDgene dataset.

Detailed parameter files for both registration methods are available via elastix.isi.uu.nl (entry par0049) and github.com/hsokooti/regun. The calculation time of ANTs was about 60 hours per registration, comparing to 12 minutes for `elastix`.

In this experiment, the evaluation is performed only on the landmarks locations, where Table 4.2 displays the distribution of reference registration errors during testing. The results of the experiments are given in Table 4.8. A scatter plot is also depicted in Fig. 4.9. Similar to the previous inter-database experiment (Table 4.7), the MAE and accuracy of the registration features are slightly better than the MAE and accuracy of the intensity-based features. However, intensity features obtained better classification

Table 4.7: Regression results for the SPREAD data trained on the DIR-Lab-4DCT data with `elastix` using 20, 100, 500 and 2000 iterations. The columns indicate the number of features (N_f), the mean absolute error (MAE), the accuracy (Acc) and the F1 score. The sub-indices c, p and w correspond to correct [0,3), poor [3,6) and wrong [6, ∞) mm classes, respectively.

	N_f	MAE	MAE$_{\mathrm{c}}$	MAE$_{\mathrm{p}}$	MAE$_{\mathrm{w}}$	Acc	F1$_{\mathrm{c}}$	F1$_{\mathrm{p}}$	F1$_{\mathrm{w}}$
Intensity	50	1.90 ± 3.63	1.56 ± 1.49	1.26 ± 1.01	10.83 ± 14.32	71.0	82.8	21.7	48.0
Registration	108	1.62 ± 3.59	1.23 ± 0.88	1.13 ± 0.81	11.53 ± 14.60	77.1	87.4	27.7	53.9
Combined	158	1.73 ± 3.56	1.36 ± 0.97	1.14 ± 0.83	11.30 ± 14.49	77.2	87.2	26.0	59.9

Table 4.8: Regression results for the DIR-Lab-COPDgene data with `elastix-advanced` and ANTs-BSplineSyN registrations trained on the SPREAD and DIR-Lab-4DCT data. The columns indicate the number of features (N_f), the mean absolute error (MAE), the accuracy (Acc) and the F1 score. The sub-indices c, p and w correspond to correct [0,3), poor [3,6) and wrong [6, ∞) mm classes, respectively.

	N_f	MAE	MAE$_{\mathrm{c}}$	MAE$_{\mathrm{p}}$	MAE$_{\mathrm{w}}$	Acc	F1$_{\mathrm{c}}$	F1$_{\mathrm{p}}$	F1$_{\mathrm{w}}$
`elastix-advanced`									
Intensity	50	2.17 ± 2.34	1.69 ± 1.35	2.81 ± 2.66	5.15 ± 4.44	64.2	77.9	20.8	64.6
Registration	108	1.84 ± 2.50	1.31 ± 1.66	2.22 ± 2.12	5.36 ± 4.29	76.0	87.6	29.9	57.6
Combined	158	1.86 ± 2.05	1.50 ± 1.16	1.92 ± 1.80	4.48 ± 4.21	75.3	86.9	29.5	66.1
ANTs-BSplineSyN									
Intensity	50	2.03 ± 2.01	1.80 ± 1.25	2.20 ± 2.27	5.30 ± 5.38	57.3	71.6	14.2	62.7
Registration	108	1.71 ± 2.39	1.43 ± 1.98	2.56 ± 2.01	5.06 ± 4.67	72.8	85.5	17.3	38.5
Combined	158	1.73 ± 1.80	1.52 ± 1.23	2.22 ± 2.27	4.45 ± 4.40	76.5	87.3	20.4	59.7

score in the wrong category. We conclude that the proposed method indeed generalizes to different settings of the same method (`elastix-advanced`), as well as registration methods with quite a different underlying transformation model (ANTs-BSplineSyN, which uses a symmetric diffeomorphic model).

4.4 Discussion

A system for quantitative error prediction of medical image registration is proposed and it is quantitatively evaluated on multiple chest CT datasets.

4.4.1 Features

In the intra-database (SPREAD) validation, it is observed that the single MIND feature can perform almost as good as the overall combined system. By adding MI and registration features, the results slightly improved. Muenzing et al. [39] did not consider MIND in their feature set and found that the most important single features in their classification experiments are mutual information and Gaussian intensity, whereas, based on Table 4.4 these features are less important than MIND in our experiments. Furthermore, the calculation time of MI for the whole image is about 3 h, as opposed to the calculation time of MIND, which is about 8 min (~3 min MIND + ~5 min pooling). Although less accurate, it is possible to reduce the calculation

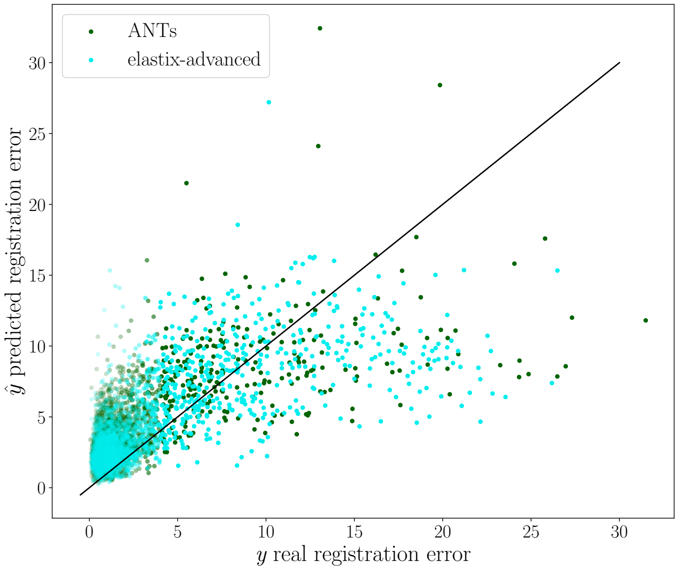


Figure 4.9: Scatter plot of real and predicted registration errors in the DIR-Lab-COPDgene database using `elastix-advanced` and ANTs-BSplineSyN registration. In total, 3000 landmarks are shown for each registration.

time of the MI feature by calculating it over a single window and then aggregate by pooling.

The modality-dependent intensity features do not increase regression accuracy on the data used in our paper. Consequently more generally applicable modality-independent features can be used, even for mono-modal problems.

Table 4.5, 4.7 and 4.8 together suggests that features in the intensity and registration categories provide complementary information, and that a better system can be obtained in terms of MAE and accuracy by considering both intensity and registration-derived features.

The intensity features were better predictors than the registration features in the intra-database experiment. However, in the inter-database experiment, the registration features outperform intensity features in terms of total accuracy and MAE. The same observation can be made for the average F1 score in the inter-database experiments using `elastix` (See Table 4.7, 4.8). For ANTs (Table 4.8), the average F1 score of the intensity-based features was slightly higher than that of the registration-based features.

The registration features contribute differently with respect to the number of iterations (See Fig. 4.8). The features std $\boldsymbol{T}^{\mathrm{L}}$ and $\mathscr{E}(\boldsymbol{T}^{\mathrm{L}})$ play important roles when the number of iterations is not enough for convergence. When the number of iterations increases, the contribution of std $\boldsymbol{T}$ and CVH go up. In the work of Muenzing et al. [39], only one registration feature, Jac, was used and they reported that the impact of this feature is relatively low in comparison with intensity-based features. We observed the same result for Jac, but it should be pointed out that the range of Jac in our database was [0.3, 3.9] so voxels with negative or very large values were not encountered.

Feature pooling improves the regression results in all evaluation measures, due to the addition of contextual information. In some features like std $\boldsymbol{T}$, average pooling contributed more to the regression performance, while in features like CVH, maximum pooling had a higher importance value (See Fig. 4.8d).

As can be seen in Table 4.6, the proposed combined system has redundant features. Hence, by removing a single feature, the system is still able to predict the registration error with almost equal MAE as the total system. However, removing these features may decrease the generalizability of the system. For example, looking at the feature $\mathscr{E}(\boldsymbol{T}^{\mathrm{L}})$ in Fig. 4.7 we see that its contribution is relatively small overall. However, in Fig. 4.8 it can be seen that while it is less important for better registration results (100, 500 and 2000 iterations), it is still important for poor registration results (20 iterations).

Considering the results in both intra and inter-database experiments (Table 4.5, 4.7 and 4.8), the conclusion to be drawn is that the proposed combined feature sets is general and robust.

4.4.2 Quantitative validation

Commonly, in image registration tasks, the distribution of registration errors is not balanced as can be seen in Table 4.3.

In the SPREAD experiment, Table 4.5 reports that in the combined experiment, the MAE of the correct and poor classes are 0.76 ± 0.65 and 1.47 ± 1.22, respectively. On the contrary, the MAE of the wrong class is 6.12 ± 5.64. It is expected that the regression error of values of the wrong class is relatively larger than that of the other classes. However, it should be emphasized that only 3.9% of samples are available to make a regression model between 6 and 81.8 mm. We tried to add more samples and make the distribution more balanced by performing registrations with different number of iterations, but there is still room for improvement for the wrong class by adding more samples and data.

In terms of classification, we obtained F1 scores of 95.4%, 38.1% and 84.4% in the classes correct, poor and wrong, respectively (Table 4.5). For the wrong class, which is arguably most important for clinical application, the precision and recall are 84.6% and 84.3%, respectively. This means that 84.6% of all samples predicted to be over 6 mm are correct and the proposed method caught 84.3% of larger registration errors. Muenzing et al. [39] obtained F1 scores of 95.3%, 73.8% and 86.6% in the classes

[0,2), [2,5) and [5, ∞) mm. They achieved a better F1 score in the poor class and they also reported zero overlap between the correct and wrong classes. However, the comparison between the two methods is not easy because of the differences in the data. For example, the slice thickness in SPREAD is 2.5 mm, while it is 1 mm for Muenzing's data, which may affect the performance especially in the poor class. Moreover, we generated the classes by thresholding the regression values. Thus, the forests are optimized for regression not for classification.

4.4.3 Qualitative validation

Muenzing et al. [32] generated an uncertainty map by spatial interpolation of landmark-based quality estimates. On the contrary, our proposed system, which is trained on landmark locations, can be applied in all regions of the image. We showed this for two example images, see Fig. 4.5. It can be easily visualized that in the blue region, images are matched correctly. On the other hand, by tracking the vessels in the red region misalignment can be seen. Another note about the prediction is that there are no abrupt changes, and error varies smoothly from blue to yellow and then red, even though the error is predicted for each voxel independently.

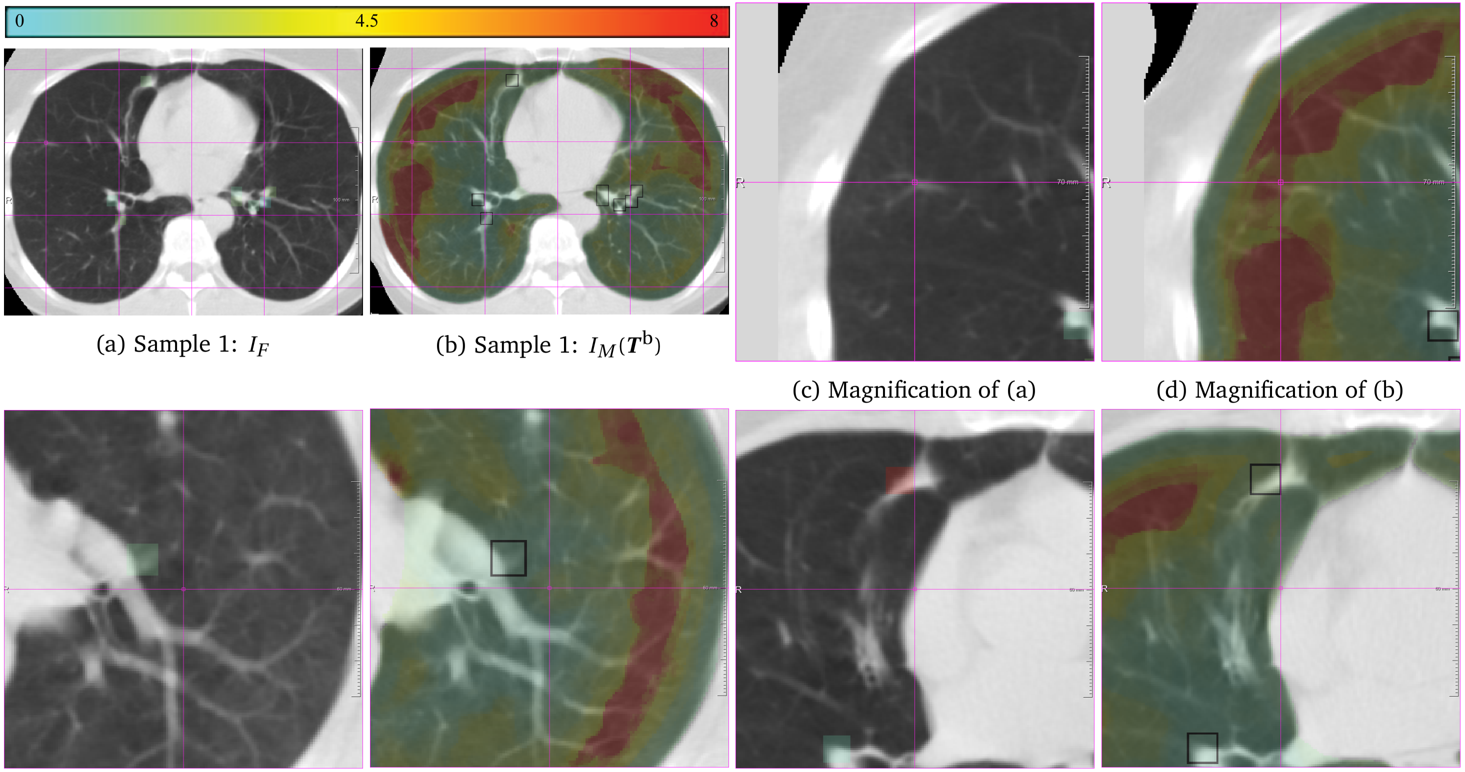
Another example is given in Fig. 4.10(a-d). Although all landmarks indicate that the registration error is small in this slice, the quantitative results found several misregistered regions, which implies that few landmarks may not be sufficient to assess the registration quality of the whole image. In Fig. 4.10(e, f), it can be observed that the performance in the homogeneous area (left side of the images) is as good as the performance in the area with structure. The main reason of acceptable performance in the homogeneous area is that the training samples consist of landmarks as well as their neighborhood region, which can be homogeneous. Thus, the system is trained both for homogeneous regions and regions with structure.

Another example is given in Fig 4.10(g, h), where the proposed system is not able to predict the registration error correctly because of a shift in the slice direction.

4.4.4 Limitations

Discrete optimization: If the optimization method is less or not dependent on the initial state, for instance for discrete optimization methods [34, 106], many of the proposed registration features, which are generated by varying the initial transformation of the registration, are not informative anymore. In such cases, instead of std $\boldsymbol{T}$ or std $\boldsymbol{T}^{\mathrm{L}}$, other measures can be used. For example, by utilizing the adaptive mean-shift algorithm, the local standard deviation of the displacement distribution can be calculated [106].

Anatomical changes: The proposed method is trained in such a way that any dissimilarity between the fixed and moving images is counted as misalignment in registration. In case of anatomical changes this assumption may be invalid, but



(a) Sample 1: I_F

(b) Sample 1: $I_M(\boldsymbol{T}^{\mathrm{b}})$

(c) Magnification of (a)

(d) Magnification of (b)

(e) Sample 2: I_F

(f) Sample 2: $I_M(\boldsymbol{T}^{\mathrm{b}})$

(g) Sample 3: I_F

(h) Sample 3: $I_M(\boldsymbol{T}^{\mathrm{b}})$

Figure 4.10: Several samples from the SPREAD dataset. The left column shows the fixed image with the ground truth registration error overlaid in color. The right column shows the moving image after registration with the registration error predicted by the proposed method overlaid in color.

typically prior knowledge of the underlying anatomy is required to determine which regions are allowed to be "misaligned" because of anatomical changes and which are not [107]. The proposed method highlights all changes, coming from misalignment or from anatomical change.

4.4.5 Future work

In the proposed method we predict the misalignment as an Euclidean distance in millimeters, rather than a 3D vector representing residual displacement. This is mostly because the features used in the system are not direction-wise, especially the local intensity features. The use of features that include directional information may help the system to be used in predicting the registration error in each direction, which is then effectively a new registration method.

The proposed method was tested on chest CT scans. Since the proposed features are generic and modality-independent, the overall method can in principle be applied to other modality data from other anatomical regions. The performance in such cases however remains to be investigated.

The uncertainty of affine registration is not measured in this work. Defining a gold standard for this mid-phase result is a complex task. However, extending the experiments to other databases where only affine transformations are applicable can be done in the future.

Instead of manually defined features, it is possible to use convolutional neural networks, which can learn features automatically. Eppenhof et al. [44] predicted the Euclidean distance of registration error. Our own work on CNNs for registration [17] can also be modified to predict registration uncertainty in a direction-wise manner. Both methods are trained only based on intensity, where the current paper shows that registration-derived information still contributes to a better regression. Thus, adding registration information to the neural networks should probably be considered as well.

A larger set of corresponding points annotated more densely throughout the scan could potentially also benefit training of the regression forest. In addition, experimenting on multi-modality data and investigating the contribution of all introduced features on them are future plans of this work.

Finally, the uncertainty map produced by the proposed method may be exploited to improve local registration results.

4.5 Conclusion

In this paper we proposed a method based on random regression forests to predict registration accuracy on chest CT scans from registration-based as well as intensity-based features. We introduced the variation in registration result from differences in initialization (std $\boldsymbol{T}$) and CVH, which showed high feature importance in several experiments. Registration-based features provided additional information on registration error with respect to intensity-based features.

The regression method was evaluated on data from the SPREAD study and predicted the registration error with a mean absolute error of 1.07 ± 1.86 mm. The proposed method gained promising results on inter-database validation with a regression error of 1.76 ± 2.59 mm.

Acknowledgments

This work is financed by the Netherlands Organization for Scientific Research (NWO), project 13351. Ben Glocker received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 757173, project MIRA, ERC-2017-STG). Dr. M.E. Bakker and J. Stolk are acknowledged for providing a ground truth for the SPREAD study data used in this paper. We would like to thank Dr. R. Castillo and T. Guerrero for providing the DIR-Lab database.

5

Hierarchical Prediction of Registration Misalignment using a Convolutional LSTM: Application to Chest CT Scans

This chapter was adapted from:

H Sokooti, S Yousefi, M Elmahdy, BP Lelieveldt, and M Staring. **Hierarchical Prediction of Registration Misalignment using a Convolutional LSTM: Application to Chest CT Scans**, *IEEE Access*, 2021.

Abstract

In this paper we propose a supervised method to predict registration misalignment using convolutional neural networks (CNNs). This task is casted to a classification problem with multiple classes of misalignment: “correct” 0-3 mm, “poor” 3-6 mm and “wrong” over 6 mm. Rather than a direct prediction, we propose a hierarchical approach, where the prediction is gradually refined from coarse to fine. Our solution is based on a convolutional Long Short-Term Memory (LSTM), using hierarchical misalignment predictions on three resolutions of the image pair, leveraging the intrinsic strengths of an LSTM for this problem. The convolutional LSTM is trained on a set of artificially generated image pairs obtained from artificial displacement vector fields (DVFs). Results on chest CT scans show that incorporating multi-resolution information, and the hierarchical use via an LSTM for this, leads to overall better F1 scores, with fewer misclassifications in a well-tuned registration setup. The final system yields an accuracy of 87.1%, and an average F1 score of 66.4% aggregated in two independent chest CT scan studies.

5.1 Introduction

Most image registration techniques do not provide insight in the local misalignment after registration. It is common to manually inspect the registration quality afterwards, which is time-consuming and prone to inter-observer errors as well as human fatigue. A fast automatic dense map indicating the misalignment locally has quite a few applications in medical imaging. This dense misalignment map can be utilized in radiation dosimetry [108], image-guided interventions [31], for improving the registration quality automatically [32] or semi-automatically [84]. Moreover, a fast automatic prediction of registration misalignment could substantially reduce the manual assessment time.

Several intensity-based and registration-based features were proposed as a surrogate for registration misalignment. Park et al. [37] proposed normalized local mutual information (NMI) and Rohde et al. [38] utilized the local gradient of the NMI as a surrogate for misregistration. Schlachter et al. [85] reported that the histogram intersection, which is a distance measure between the histogram of intensities of a pair of images [109], performs well as a visual assistant to a human expert in detecting local registration quality. Although the mentioned metrics can represent the registration error, it has been shown by Rohlfing [110] that image similarities cannot necessarily distinguish accurate from inaccurate registrations. Hub et al. [35] proposed performing multiple registrations with perturbations in the B-spline grid [6] as a measure of registration uncertainty. Kybic [94] proposed bootstrapping over pixels in the cost functions. Other approaches like block matching [111] and polynomial chaos expansions [112] are utilized in the context of detecting registration misalignment. However, these algorithms are very time-consuming.

In probabilistic image registration, an uncertainty map can be provided after the registration [34, 96, 106]. This uncertainty map commonly is counted as a surrogate for image registration error. However, Luo et al. [113] reported that the uncertainty derived from probabilistic image registrations might not necessarily correlate with the registration error.

Several machine learning approaches have been used in assessing the registration quality. Muenzing et al. [39] cast the problem to a classification task. They extracted several intensity-based features around a number of distinctive landmarks in chest CT images. Sokooti et al. [55, 33] extracted both intensity and registration-based features around a dilated region of landmarks and trained a regression forest to predict the registration error. Drawbacks of these methods are that training is based on a limited number of manual landmarks, and/or can only be applied to nonrigid registration.

Deep learning-based methods have been presented recently and achieved promising results for medical image registration [17, 23, 114]. Predicting the registration error with a CNN-based approach was recently proposed by Eppenhof et al. [40]. They used a single scale method and predicted registration misalignment smaller than 4 mm. Senneville et al. [115] proposed a deep learning method to classify brain MR registrations as usable or non-usable. This method cannot predict misalignment locally, for nonrigid image registration.

Hierarchical approaches have been used in many tasks in the field of image classification. Salakhutdinov et al. [116] proposed a hierarchical classification model, in which objects with fewer occurrences can borrow statistical strength from related objects that have many training examples. Ristin et al. [117] reported that taking into account the hierarchical relations between categories and subcategories can improve the performance of classification. Such an approach has also been used in recent deep learning methods. Redmon et al. [118] in their proposed method for object detection, YOLO9000, predict labels in a hierarchical approach using conditional probability. Chen et al. [119] predict abnormality labels in chest X-ray images using a similar hierarchical approach with conditional probability. They added another stage with unconditional probabilities and reported better performance in comparison with only a single stage with conditional probability. Taherkhani et al. [120] reported that utilizing coarse images can improve weakly supervised fine image classification performance. Guo et al. [121] reported that utilizing a convolutional LSTM [122] and predicting the labels from coarse to fine, can improve the accuracy of the classification of both coarse and fine labels. In their method, the CNN and LSTM extract discriminative features and jointly optimize the fine and coarse labels classification. A similar hierarchical LSTM approach has been utilized in music genre classification [123]. In the aforementioned methods, the hierarchical approach is only applied on the network outputs (coarse and fine labels), while the inputs are kept similar in all steps of the hierarchy.

In this work, inspired by the hierarchical classification idea of [121], we propose a hierarchical convolutional LSTM approach to densely predict the registration misalignment. Moreover, we incorporate multi-resolution information for the inputs as well as the outputs. This way, the LSTM takes input images from coarse to fine resolution and progressively predicts output labels from coarse to fine. We propose to use a pre-trained registration network to encode the input image pair in a latent space, and utilize an LSTM decoder to predict the final labels from this latent space. We trained our deep learning model on image pairs artificially generated from real data, as a data augmentation step. In this way, in contrast to [39] and [33], we have access to many training samples instead of a small number of manually annotated landmarks. Different from earlier deep learning methods, the proposed method can be used to predict the registration error for any registration paradigm, including rigid and nonrigid registration. Different from [40], the proposed method is capable of detecting relatively large registration misalignments. The inference time of the

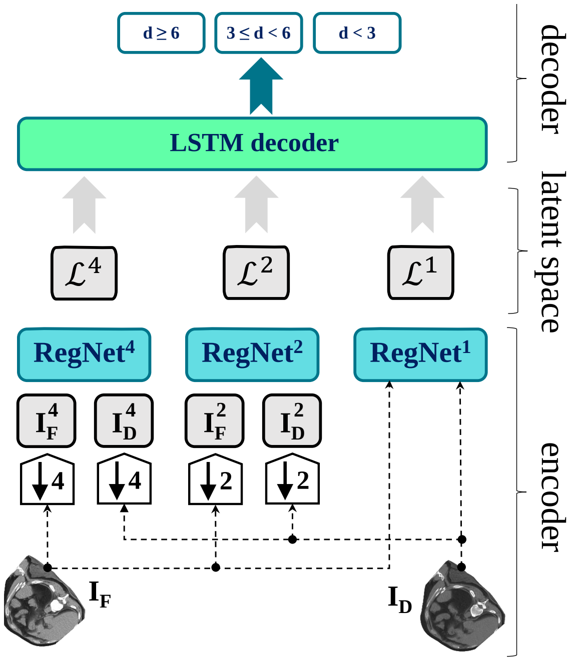


Figure 5.1: Block diagram of the proposed system. In the encoder, a pair of images is given as the input. Three RegNet architectures [18] process the input images over three resolutions (↓4, ↓2, 1) and generate a latent representation (the encoded feature maps $\mathscr{L}^i$) for each resolution. All RegNet blocks are architecturally identical, but are initialized with weights from pre-trained networks on different resolutions. In the LSTM decoder, the latent representations $\mathscr{L}^i$ are decoded to labels corresponding to the local misalignment class d.

proposed method is approximately 2.8 seconds on a 3D patch of size 205 × 205 × 205, which is substantially faster than methods involving multiple registrations like [94, 35, 33].

In Section 5.2, we introduce the network architectures (5.2.1) and explain the training data generation process (5.2.2). In Section 5.3, we describe the data sets used in this study (5.3.1), the detailed setup of the experiments (5.3.2), and the evaluation measures (5.3.3). The tuning of hyper-parameters (5.3.4) and the results 5.3.5, 5.3.6) are reported afterwards. Finally, the Discussion (Section 5.4) and Conclusion (Section 5.5) are presented.

5.2 Methods

A general block diagram of the proposed method is shown in Fig. 5.1. The input of the network is a pair of images consisting of a fixed image I_F and a deformed moving image I_D, resulting from an arbitrary registration method. The input image pair is then downsampled and encoded by a deep learning registration network at three resolutions. The latent representations $\mathscr{L}^i$ are subsequently fed to a decoder (an LSTM), where the decoder predicts misregistration labels d for each voxel, corresponding to the local misalignment. The LSTM not only considers the encodings at the three resolutions, but also considers these in a coarse-to-fine, hierarchical manner.

5.2.1 Network architectures

5.2.1.1 Encoder

In the encoder, an image pair (I_F, I_D) is encoded to create a latent representation of the input pair and their spatial relation. Such an encoder may be trained from scratch, or a pre-trained architecture can be chosen. Popular examples of the latter is to use a VGG or a ResNet network trained on large-scale natural images [124, 125], sometimes also used to compute a perceptual loss in a downstream task [126]. A downside of such an approach is that each of the input images is encoded separately, and subsequently the spatial relation between the input images is not represented. In addition, as reported by Raghu et al. [127], for medical imaging tasks a network trained on similar data is favored over a network trained on natural images. Instead, we therefore propose to encode the input pair by a pre-trained medical image registration network, thus allowing the direct encoding of a pair of images, while also representing the spatial relation between them.

Any registration network from the literature can be used here, and we opt for the RegNet architecture [18, 17], which we previously proposed for the registration of chest CT scans. Since this network achieved promising results, it is potentially a good candidate for the task of predicting registration misalignment as well. The RegNet architecture is given in Fig. 5.2. This design is identical to the U-Net-advanced (Uadv) design proposed in [18]. The last three layers from the original design are excluded here, and the high dimensional feature maps from the now last layer are used as a latent representation of the input pair, and thus as input for the decoder. As illustrated in Fig. 5.1, we utilize three separate encoders, each receives an input image pair at a different resolution, using a down-sampling factor of four (↓4), two (↓2) and 1 (i.e. the original resolution). This way latent representations are built at three different scales.

The RegNet architecture is a patch-based design where the size of the inputs and output are $101 \times 101 \times 101$ and $25 \times 25 \times 25$, respectively. All convolutional layers use batch normalization [70] and ReLu activation [71], except for the trilinear upsampling layer, in which a constant trilinear kernel is used. The total number of parameters in this design is 737,430.

The weights of the three encoders are initialized with the pre-trained RegNet[i]

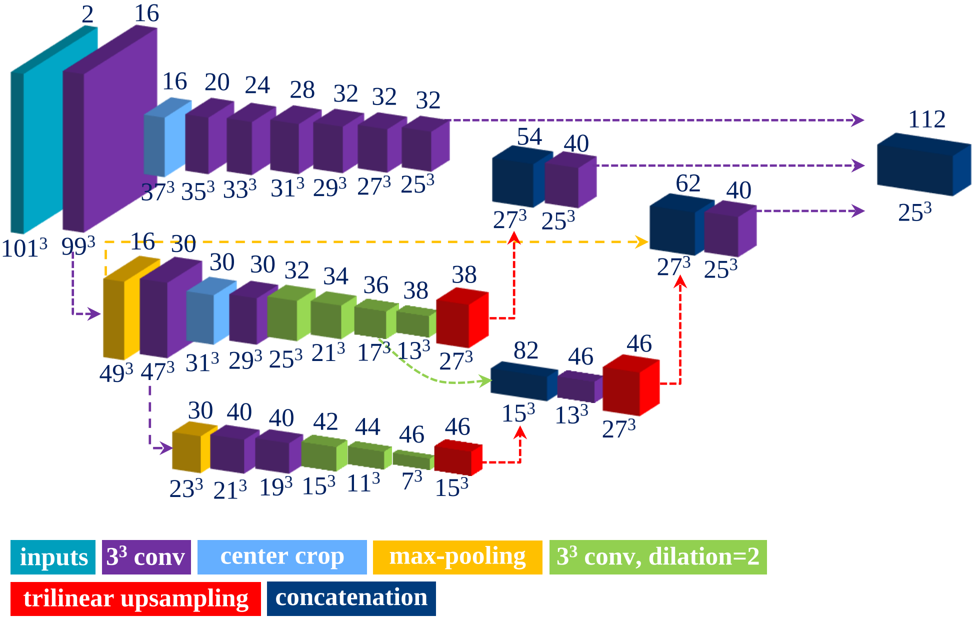
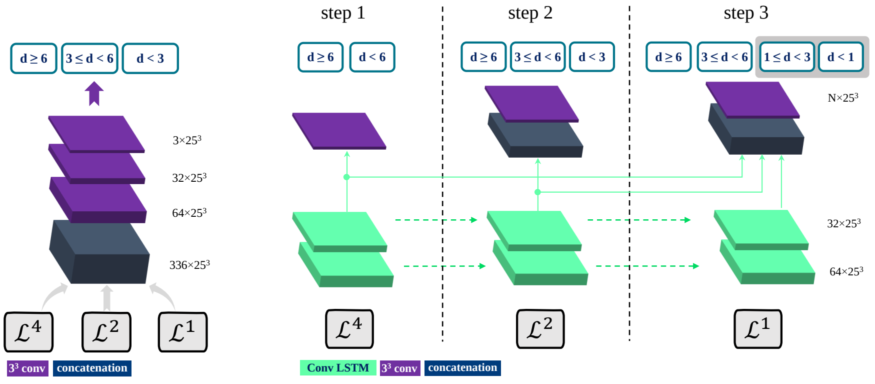


Figure 5.2: The RegNet architecture used for encoding the input image pair. This architecture is identical to the U-Net-advanced (Uadv) design proposed in [18], with the last three layers excluded. The number of feature maps and the spatial size are shown on top and bottom of each layer, respectively.

networks (see Fig. 5.1), that were previously trained for image registration [18]. Below, we report experiments both with freezing these weights and with keeping them trainable. When keeping them trainable, all layers are kept trainable, as recommended by Tajbakhsh et al. [128].

5.2.1.2 Decoder

In the decoder, the latent representations at each of the three resolutions $\mathscr{L}^i$ are considered to predict three output labels corresponding to registration misalignment: correct [0,3) mm, poor [3,6) mm and wrong $[6,\infty)$ mm [33]. A straightforward choice for the decoder is to concatenate the latent feature maps and feed them to a convolutional neural network to predict the final labels. This approach is illustrated in Fig. 5.3a and is named multi-scale CNN. Instead, we propose a hierarchical approach using convolutional LSTM (Long Short-Term Memory) layers similar to [121] as they reported that predicting the labels from coarse to fine can improve the overall accuracy of the classification of fine labels in natural images. The coarse labels usually share a set of global features and for the fine labels more distinctive local properties are



(a) Multi-scale CNN decoder

(b) Hierarchical LSTM decoder

Figure 5.3: The decoder. The latent representations $\mathscr{L}^i$ of the three resolutions ↓4, ↓2 and 1 are merged and the final output predicts three misalignment labels: correct [0,3) mm, poor [3,6) mm and wrong [6, ∞) mm. In the CNN decoder (a), merging is done using concatenation. In the LSTM decoder (b), the latent representations $\mathscr{L}^i$ are given in sequence and the misalignment labels are gradually refined in a hierarchical manner. The labels inside the shaded boxes in the top-right of the figure represent the auxiliary labels.

extracted.

The LSTM unit was first proposed for machine translation where the input, output, and hidden states are all modeled as temporal sequences using fully connected units [129]. As this approach does not capture the spatial relations in the data, Shi et al. [122] proposed a convolutional LSTM unit, where the fully connected (FC) layers are replaced by convolutional layers. This way the unit is capable of capturing and encoding spatio-temporal information for visual series. We can imagine inputs and state as vectors standing on a spatial grid. The future state of a cell in the grid is calculated by the inputs and past states of its neighbors.

In the proposed LSTM decoder (Fig. 5.3b), rather than supplying the three latent representations $\mathscr{L}^i$ all at once, they are provided in sequence. Starting with $\mathscr{L}^4$, a coarse prediction of the registration error is first made, predicting only two labels: 'good' registration with an error in the range $[0, \theta_1)$ mm, and 'bad' registration with an error higher than that i.e. $[\theta_1, \infty)$ mm. In the experiments for example we have used $\theta_1 = 6$ mm. In the next time step of the convolutional LSTM, the $\mathscr{L}^2$ features are additionally considered, combining them with the hidden state of the previous time step. Now the output predictions are refined into three classes $[0, \theta_2)$ mm, $[\theta_2, \theta_1)$ mm and $[\theta_1, \infty)$ mm. We keep all the output probabilities unconditional similar to [121]. In the last time step, the latent representation $\mathscr{L}^1$ is used and combined with the hidden state, further refining the output prediction with splitting the previous smallest class to $[0, \theta_3)$ mm and $[\theta_3, \theta_2)$ mm. This way the predictions are built up in a hierarchical manner, step-by-step incorporating the multi-resolution embeddings of the input pair and step-by-step refining the registration error prediction.

In the final convolutional layers of both decoder designs, the softmax activation is used. For other convolutional layers in the CNN-based decoder, batch normalization and ReLu activation are utilized. In the LSTM design, cell outputs, hidden states, and gates (input, forget, output) have similar settings as in [122]. An additional output is allocated for each coarse label. For instance, in Fig. 5.3b, six outputs are available, four of them for fine labels and two for coarse labels. We perform experiments for various values of θ_i, where $i \in \{1,2,3\}$ and $\theta_1 \geq \theta_2 \geq \theta_3$.

5.2.2 Training data generation

In order to train the networks, we propose to artificially generate image pairs from the available real data. The main advantage of artificial generation is that numerous number of training samples can be obtained in an inexpensive way. Moreover, a dense ground truth is made, which is not achievable with other forms of ground truth such as manual landmarks or segmentation maps.

We use a similar approach as in [18] to artificially generate the DVFs and deformed image. Four types of artificial deformation are applied:

single frequency: This type of DVF is generated by perturbing B-spline grids. Since the grid knots are uniformly spaced, the generated DVF has only one random spatial frequency.

mixed frequency: A combination of the single frequency DVF filtered by a Gaussian kernel with a smaller sigma.

respiratory motion: Simulating the respiratory motion by expansion of the chest in the transversal plane, transition of the diaphragm in craniocaudal direction [35]. Finally, a random "single frequency" deformation is added.

identity transform: This type represents no misalignment between the images.

After creating the deformed images with the generated DVFs, to make the deformed images more realistic, several intensity augmentations are performed:

Gaussian noise: Gaussian noise with a standard deviation of $\sigma_N = 5$ is added to the deformed image.

Sponge model: Multiplying the intensity of the deformed moving image by the inverse of the determinant of the Jacobian of the transformation. This is an approximation based on the theory of mass preservation in the lung during breathing [73].

By applying the proposed artificial DVF generations, many image pairs can be generated for each image, by varying the hyper-parameters corresponding to each category.

5.3 Experiments and Results

5.3.1 Data

Experiments are performed using three chest CT studies: The DIR-Lab-COPDgene [75], the DIR-Lab-4DCT [74] and the SPREAD [47] studies.

In the DIR-Lab-COPDgene study, ten cases are available in inhale and exhale phases. The average image size and the average voxel size are $512 \times 512 \times 120$ and $0.64 \times 0.64 \times 2.50$ mm, respectively. 300 corresponding landmarks are manually annotated in each case.

In the DIR-Lab-4DCT study, ten cases with varying respiratory phases are available. We selected the maximum inhalation and maximum exhalation phases, as more manual landmarks are available in these phases (300 landmarks). The size of the images is approximately $256 \times 256 \times 103$ with an average voxel size of $1.10 \times 1.10 \times 2.50$ mm.

In the SPREAD study, 21 cases are available. Each case consists of a baseline and a follow-up image, in which the follow-up is taken after about 30 months. Both baseline and follow-up are acquired in the maximum inhale phase. The size of the images is about $446 \times 315 \times 129$ with a mean voxel size of $0.78 \times 0.78 \times 2.50$ mm. About 100 well-distributed corresponding landmarks were previously selected [73] semi-automatically on distinctive locations [48]. Two cases (12 and 19) are excluded because of the high uncertainty in the landmark annotations [73].

5.3.2 Experimental setup

5.3.2.1 Training data

In the SPREAD study, 10 , 1, and 8 cases are used for the training, validation, and test sets, respectively. The DIR-Lab-COPD study is used for training and validation only, where 9 cases are used for training and the remaining case for validation. The entire DIR-Lab-4DCT database (10 cases) is used as an independent test set. The validation set is mainly used for tuning the hyper-parameters and selecting the best approach. Since we initialized the weights of RegNet from the study of [18], we kept the training, validation, and test sets identical to that study, to avoid data leakage.

To generate training pairs, we use the artificial generations introduced in 5.2.2. The maximum magnitude of the DVF in each axis is set to 10 mm, so the maximum vector magnitude is about 17 mm. For each single image, 28 artificial DVFs and deformed images are generated by assigning random values to the variables of the single frequency, the mixed frequency and the respiratory motion deformations. Thus, in the training phase, a total number of 1064 artificially generated image pairs are used. All images are resampled to an isotropic voxel size of $1.0 \times 1.0 \times 1.0$ mm.

In the training phase, the patches are balanced based on the magnitude of the artificial DVFs. The probabilities of selecting patches in the range [0, 3), [3, 6) and 6, ∞) mm are 60%, 20% and 20%, respectively. This balancing is performed to make the training set more similar to the real world scenarios as the distribution of landmarks in the first range is usually higher.

5.3.2.2 Real image pairs

In this experiment, we estimate the registration error after registration in cases from the test set and compare it with the ground truth landmarks. Both fixed and moving images are taken from the same patient at different time points. In order to create a generic evaluation study, we collect samples by performing affine and four various conventional nonrigid registrations using 20, 100, 500, and 2000 iterations corresponding to overall poor registration quality to overall high quality registration. The common registration settings are: metric: mutual information, optimizer: adaptive stochastic gradient descent, transform: B-spline ([6]), number of resolutions: 3. After performing registration on the original fixed and moving images, the fixed and the deformed moving image after the registration are given as inputs to the proposed misalignment estimation method.

We define the target registration error (TRE) as the Euclidean distance after registration between the corresponding ith landmarks:

$$\mathrm{TRE}^{\mathrm{i}} = \|\boldsymbol{x}_F^{\boldsymbol{i}} - \boldsymbol{x}_D^{\boldsymbol{i}}\|_2, \tag{5.1}$$

where $\boldsymbol{x}_F$ and $\boldsymbol{x}_D$ are the corresponding landmark locations on the fixed and deformed moving images, respectively. A misalignment label is then assigned to each landmark, based on the magnitude of the TRE. The misalignment labels are defined based on the TRE value.

5.3.2.3 Network optimization

Optimizing the neural networks is done by the Adam optimizer [130] with a constant learning rate of 0.001. A stochastic mini-batch method is used with a batch size of 10. The cross-entropy loss is used for all experiments. In the LSTM design, the cross-entropy loss is applied to unconditional probabilities for all steps similar to [121]. The loss function is defined as follows:

$$\mathrm{loss} = -\frac{1}{N}\sum_{i=1}^{N}\Big(\sum_{s=1}^{S}\sum_{c\in C^s} 1\{x_i^s = c\}\log p_c\Big), \tag{5.2}$$

where N is the total number of voxels in a mini-batch, S denotes the number of steps, C^s represents the classes at step s, and p_c is the probability of class c in the output. The training is performed for 30 epochs by an NVidia RTX6000 with 24GB memory.

5.3.2.4 Software

The convolutional neural networks are implemented in Tensorflow [76], and image handling and artificial training data generation is implemented with SimpleITK [51]. `elastix` [52] is used to perform the conventional image registrations.

5.3.2.5 Additional methods

For further comparisons, two additional CNN methods are added: single-scale CNN and RegNet-t. In the single-scale CNN, only the encoded feature maps of the original resolution $\mathscr{L}^1$ is used. The weights of the encoder are kept trainable similar to the multi-scale CNN. In the RegNet-t experiment, first a three-resolution registration is performed by RegNet over the input pair [18]. The registration is performed over scales four, two and one in sequence, in which the input of each resolution is the fixed and deformed moving image of the previous resolution. Then, the magnitude of the predicted displacement vector field (DVF) is calculated and thresholded in the following ranges: [0,3), [3,6) and [6, ∞) mm. Finally, the labels "correct", "poor" and "wrong" are assigned to them, respectively.

In addition, the proposed multi-stage hierarchical LSTM design is compared to a conventional learning-based method using random forests (RF), published earlier [33]. The random forests were trained on several hand-crafted intensity-based and registration-based features extracted from landmark neighborhoods. The output of the random forests predicted the registration error in mm. Three classes were generated by quantizing the regression results within the ranges [0,3), [3,6), and [6, ∞) mm, similar to the current study.

5.3.3 Evaluation measures

All evaluations are computed only from the landmark locations to maximize the quality of the ground truth. The misalignment labels are defined as correct, poor and wrong, when the TRE is in range [0,3), [3,6) and [6, ∞) mm, respectively, similar to [33]. We report the following statistics: overall accuracy, F1 score for each label separately, the average $\overline{\text{F1}}$ of the separate F1 scores, the number of misclassifications between the wrong and the correct label (two categories apart called cw misclassification), and finally Cohen's kappa coefficient (κ) of the confusion matrix. The accuracy may be biased to the labels with a higher number of samples, whereas the $\overline{\text{F1}}$ and κ coefficient are more robust for imbalanced distributions.

Table 5.1: Landmark-based results on the training and validation set for tuning hyper-parameters. We report the mean values over all five registration settings: affine and B-spline registration after affine with 20, 100, 500, and 2000 iterations. The sub-indices c, p, and w correspond to the correct [0,3), poor [3,6), and wrong [6, ∞) mm classes. The best method is shown in bold and the second best method is shown in green. Total number of landmarks for all five registrations in SPREAD (cases 1 to 11) and DIR-Lab COPDgene studies are 5455 and 15000, respectively

		SPREAD (case 1 to 11)							DIR-Lab COPDgene (case 1 to 10)						
encoder	decoder	$F1_c$	$F1_p$	$F1_w$	$\overline{F1}$	Acc	κ	cw misclass	$F1_c$	$F1_p$	$F1_w$	$\overline{F1}$	Acc	κ	cw misclass
frozen	multi-scale CNN	85.8	52.5	83.5	73.9	78.1	0.58	39	**77.6**	52.5	85.8	72.0	**77.0**	0.60	387
trainable	multi-scale CNN	90.0	62.5	82.3	78.3	83.1	0.66	32	72.2	**61.4**	85.1	72.9	75.8	0.60	209
frozen	LSTM 6-3-1	92.4	54.9	83.3	76.9	85.5	0.68	52	76.9	38.5	**86.9**	67.4	76.2	0.59	391
trainable	LSTM 6-3-1	**93.0**	**63.6**	82.3	**79.6**	**86.3**	**0.71**	25	74.6	59.4	85.6	**73.2**	76.1	**0.61**	**148**
trainable	LSTM 12-6-3	83.0	54.2	**84.5**	73.9	75.6	0.56	**15**	56.7	56.3	84.6	65.8	71.9	0.53	368
trainable	LSTM 6-3-3	88.7	58.9	83.6	77.1	81.5	0.64	28	60.6	56.4	84.2	67.1	71.9	0.53	253

5.3.4 Results on the validation set

This experiment is mainly designed for tuning the hyper-parameters, i.e. the splitting values for the LSTM and to choose between the trainable and the frozen weights approach. We experiment with the two decoder architectures introduced in Section 5.2.1.2: the multi-scale CNN decoder and the hierarchical LSTM decoder. The encoding architecture is kept identical in all experiments and all weights are initialized from the pre-trained RegNet [18]. The results are reported for both frozen and trainable encoder weights. In the trainable experiment, the weights of all layers are kept trainable. Additionally, three different splitting values for the LSTM designs are tested as well.

Table 5.1 gives the results on the training and validation sets for the decoders with similar encoder design with frozen and trainable approaches. Please note that the training was performed on the artificial image pairs. However, these results are reported over real images pairs on the landmark locations. Total number of landmarks for all five registrations in SPREAD (cases 1 to 11) and DIR-Lab COPDgene studies are 5455 and 15000, respectively.

First, we compare the encoding parts between frozen and trainable approaches. In this evaluation, the splitting values of the LSTM design are set to 6, 3, 1 for θ_1, θ_2 and θ_3, respectively. As is shown in the top four rows of Table 5.1, based on $\overline{F1}$, κ coefficient and the number of misclassifications between the wrong and the correct label (cw misclass), a consistent improvement can be achieved by utilizing a trainable encoder. The improvement of $\overline{F1}$ in the SPREAD study is from 73.9% to 78.3% and 76.9% to 79.6%, and in the DIR-Lab COPDgene study from 72.0% to 72.9% and 67.4% to 73.2% for the multi-scale CNN and the hierarchical LSTM architecture, respectively. Accuracy (Acc) is more biased towards category c, as the number of samples for this label is much higher than for the other labels. In the SPREAD dataset, $F1_c$ and the accuracy of the trainable encoders are better. However, in the DIR-Lab COPDgene set, $\mathrm{F1_c}$ and the accuracy of the frozen encoders are slightly better. On the other hand, the number of outliers significantly decreases in the DIR-Lab COPDgene study. All in all, we select the trainable approach for the encoder in the remainder of the paper.

Comparing the two decoders (with trainable encoder), the LSTM design obtained better performance in terms of $\overline{\mathrm{F1}}$, κ coefficient, the number of outliers, and accuracy, compared to the CNN, on both datasets. We keep both designs for further experiments on the independent test data.

We additionally experiment with the hierarchical splitting approach of the LSTM design, using various splitting values θ_i: 6-3-1, 12-6-3 and 6-3-3. We keep the misalignment labels of the last step equal to [0, 3), [3, 6) and [6, ∞) mm by merging the auxiliary labels. Therefore, in the LSTM design with the 6-3-1 splitting approach, labels [0, 1), [1, 3) are merged into a single label [0, 3), and in the LSTM design with the 12-6-3 splitting approach, labels [6, 12), [12, ∞) are merged into a single label [6, ∞). The results are given in the bottom two rows in Table 5.1. Based on the $\overline{\mathrm{F1}}$, κ coefficient and the number of cw misclassifications, the hierarchical splitting with values 6-3-1 achieved better performance. The $\mathrm{F1_w}$ score of LSTM 12-6-3 in the SPREAD study are relatively high. On the other hand, the $\mathrm{F1_c}$ of LSTM 6-3-1 is higher than the other LSTM designs. This indicates that utilizing an auxiliary label in a specific range can improve the performance in that range. All in all, we select the LSTM with 6-3-1 splitting values for the remainder of the paper.

5.3.5 Results on the independent test set

In this section, we investigate the performance of the proposed decoders in unseen test sets, i.e. the SPREAD study cases 13 to 21 and the DIR-Lab 4DCT cases 1 to 10. The total number of landmarks for each registration in SPREAD (case 13 to 21) and DIR-Lab 4DCT studies are 783 and 3000, respectively. For further comparisons, two additional methods are added in this experiment: single-scale CNN and RegNet-t (see Section 5.3.2.5). The landmark-based results are reported in Table 5.2 within five various registration settings (similar to the validation experiment): affine transformation, B-spline transformation with 20, 100, 500, and 2000 iterations. The B-spline registrations are performed after the initial affine transformation. The aggregation of all five registrations are presented in the “total” row.

As seen in Table 5.2, among the classification networks, in the “total” row, the multi-scale CNN and LSTM 6-3-1 achieved better results in terms of $\overline{\mathrm{F1}}$ score and the number of cw misclassifications. This demonstrates that utilizing information from different scales can improve the performance. The LSTM design performed better in the SPREAD study based on all of the measures in this table $\mathrm{F1_c}$, $\mathrm{F1_p}$, $\mathrm{F1_w}$, $\overline{\mathrm{F1}}$, accuracy (Acc), κ coefficient and the number of cw misclassifications. In the same evaluation in the DIR-Lab 4DCT study, there is no consistent superiority among the multi-scale classification networks. In terms of $\overline{\text{F1}}$, the multi-scale CNN gained slightly better results i.e. 75.9% in comparison with single-scale CNN (73.9%) and LSTM (73.1%). All in all, based on the number of cw misclassifications, the multi-scale CNN and the LSTM design performs better than the single-scale CNN.

Strikingly, direct quantization of the RegNet encoder (method RegNet-t) performs quite well for affine registration and for coarse B-spline registration with a small number of iterations (20 and 100), leading to improved kappa values compared to the other three classification networks. For instance, for affine registration, RegNet-t achieved the highest $\overline{\text{F1}}$ score of 78.2% and 83.4% for SPREAD and DIR-Lab 4DCT, respectively. However, for more realistic B-spline registration with a larger number of iterations, the LSTM and the multi-scale CNN methods perform better. For example for B-spline registration with 2000 iterations, a $\overline{\text{F1}}$ score of 68.9% and 63.9% were obtained for the LSTM on the SPREAD and DIR-Lab 4DCT datasets, respectively. Notably, the LSTM decoder performs much better in terms of the number of cw misclassifications compared to RegNet-t, especially for the DIR-Lab 4DCT dataset where this number decreases from 197 to 77 in the "total" row. The inference time on a 3D patch of size $205 \times 205 \times 205$ was approximately 2.4, 0.7, 1.3, and 2.8 seconds for RegNet-t, single-scale CNN, multi-scale CNN, and LSTM, respectively.

Detailed results for the LSTM 6-3-1 decoder are reported in Tables 5.3 and 5.4. Table 5.3 shows the confusion matrix for the three classes correct, poor, and wrong, for the results aggregated over all registration settings (the "total" row in Table 5.2). The vast majority of misclassifications is one category off, with only 0.23% (9/3915) and 0.51% (77/15000) of the misclassifications two categories off, for the SPREAD (case 13 to 21) and DIR-Lab 4DCT studies, respectively. The intermediate hierarchical prediction results for each of the LSTM time steps are given in Table 5.4. Such results are not available for the CNN-based decoder, as that architecture lacks the possibility for gradual refinement. In step 1, only low resolution latent representations are available ($\mathscr{L}^4$), with a prediction in two classes only: [0, 6) mm and above 6 mm. This results in F1 scores of 92.4% and 60.1% for these two classes, for the SPREAD data. The results are gradually refined, by adding higher resolution representations and by predicting more fine-grained registration error classes, see Table 5.4. It can be seen that as the LSTM refines its results, the F1_p and F1_w scores are gradually improved in both studies. From step2 to step3-merged all F1 measures improve, in particular for the DIR-Lab 4DCT study.

Visual examples of the predictions for LSTM 6-3-1, single CNN, multi CNN, and RegNet-t are illustrated in Fig. 5.4. The ground truth misalignment on the landmark locations are dilated for better visualization. The color bar in the top center image indicates the target registration error. For all predictions, a three-label output is

Table 5.2: Landmark-based results on the test set. We report metrics over all five registration settings: affine and B-spline registration after affine with 20, 100, 500, and 2000 iterations. The sub-indices c, p and w correspond to the correct [0,3), poor [3,6) and wrong [6, ∞) mm classes. The best method is shown in bold and the second best method is shown in green. Total number of landmarks for each registration in SPREAD (cases 13 to 21) and DIR-Lab 4DCT studies are 783 and 3000, respectively.

		SPREAD (case 13 to 21)							DIR-Lab 4DCT (case 1 to 10)						
registration	decoder	$\mathrm{F1_c}$	$\mathrm{F1_p}$	$\mathrm{F1_w}$	$\overline{\mathrm{F1}}$	Acc	κ	cw misclass	$\mathrm{F1_c}$	$\mathrm{F1_p}$	$\mathrm{F1_w}$	$\overline{\mathrm{F1}}$	Acc	κ	cw misclass
Affine	RegNet-t	**70.5**	**72.1**	**92.1**	**78.2**	**83.9**	**0.70**	**0**	88.4	**70.2**	**91.6**	**83.4**	**85.7**	**0.77**	12
	single CNN	41.3	51.5	87.8	60.2	75.1	0.49	7	86.1	59.8	90.7	78.9	83.1	0.72	9
	multi CNN	47.8	71.2	92.1	70.3	81.6	0.66	**0**	88.5	66.6	85.6	80.2	81.0	0.71	**2**
	LSTM 6-3-1	65.5	67.1	91.0	74.5	81.5	0.66	**0**	**88.6**	58.4	79.3	75.4	76.1	0.64	9
B-spline 20	RegNet-t	**89.6**	**67.7**	**82.8**	**80.0**	**83.0**	**0.69**	**2**	**92.0**	**67.3**	**88.7**	**82.7**	**85.5**	**0.77**	15
	single CNN	77.1	47.1	65.5	63.2	66.3	0.47	37	89.7	56.7	87.4	77.9	82.4	0.73	29
	multi CNN	83.7	64.4	82.1	76.7	77.4	0.62	**2**	90.2	64.0	82.2	78.8	80.5	0.71	**6**
	LSTM 6-3-1	88.4	65.6	82.1	78.7	81.2	0.67	**2**	91.2	57.3	77.8	75.4	78.5	0.67	**6**
B-spline 100	RegNet-t	95.0	51.4	75.3	73.9	90.0	0.60	8	**92.7**	**61.7**	**84.8**	**79.8**	**85.0**	**0.74**	25
	single CNN	84.6	30.6	53.0	56.0	73.2	0.36	42	88.6	47.4	83.9	73.3	80.1	0.67	55
	multi CNN	91.8	48.8	**76.4**	72.3	85.1	0.55	8	91.0	57.3	73.7	74.0	78.9	0.66	**9**
	LSTM 6-3-1	**95.6**	**56.1**	75.6	**75.8**	**90.4**	**0.65**	**3**	92.3	54.0	71.1	72.5	79.2	0.65	17
B-spline 500	RegNet-t	**96.7**	**48.5**	68.2	71.1	**92.7**	**0.58**	4	93.3	**55.5**	65.7	**71.5**	**82.8**	**0.64**	56
	single CNN	86.5	25.1	43.0	51.5	76.0	0.30	51	88.7	36.4	**75.4**	66.8	77.6	0.59	81
	multi CNN	93.7	43.8	73.8	70.5	88.3	0.52	10	91.4	53.3	62.8	69.2	79.0	0.61	**17**
	LSTM 6-3-1	95.7	44.4	**83.0**	**74.4**	91.4	0.57	**2**	**93.3**	50.8	60.8	68.3	81.1	0.61	23
B-spline 2000	RegNet-t	**96.7**	27.3	56.2	60.1	**93.0**	0.41	7	93.2	46.3	43.6	61.0	81.7	0.54	89
	single CNN	86.9	16.4	41.1	48.1	76.6	0.25	50	89.3	35.6	**71.7**	65.5	79.0	0.57	127
	multi CNN	93.6	24.1	72.7	63.5	87.7	0.39	8	92.8	**50.1**	57.6	**66.8**	81.2	**0.59**	41
	LSTM 6-3-1	96.2	**30.6**	**80.0**	**68.9**	92.2	**0.50**	**2**	**93.6**	42.9	55.3	63.9	**81.9**	0.56	**22**
total	RegNet-t	**94.2**	**63.4**	**87.9**	**81.8**	**88.5**	**0.76**	21	**92.4**	**61.6**	83.2	**79.1**	**84.1**	**0.73**	197
	single CNN	83.3	39.1	74.6	65.7	73.4	0.54	187	88.7	48.7	**84.4**	73.9	80.4	0.68	301
	multi CNN	90.3	59.2	87.7	79.1	84.0	0.70	28	91.2	59.2	77.3	75.9	80.1	0.68	**75**
	LSTM 6-3-1	93.6	60.4	87.8	80.6	87.4	0.75	**9**	92.3	53.8	73.2	73.1	79.4	0.66	77

Table 5.3: Confusion matrix of the landmark-based results on the test set, for the trainable LSTM 6-3-1 decoder. We report the aggregated values over all five registration settings: affine and B-spline registration after affine with 20, 100, 500, and 2000 iterations. The sub-indices c, p and w correspond to correct [0,3), poor [3,6) and wrong [6, ∞) mm classes. P and A refer to the predicted and actual labels for each class. Total number of landmarks for all five registrations in SPREAD (case 13 to 21) and DIR-Lab 4DCT studies are 3915 and 15000, respectively.

SPREAD (case 13 to 21)				DIR-Lab 4DCT (case 1 to 10)			
	$\mathrm{A_c}$	$\mathrm{A_p}$	$\mathrm{A_w}$		$\mathrm{A_c}$	$\mathrm{A_p}$	$\mathrm{A_w}$
$\mathrm{P_c}$	2441	117	3	$\mathrm{P_c}$	7526	680	70
$\mathrm{P_p}$	209	371	72	$\mathrm{P_p}$	492	1757	1656
$\mathrm{P_w}$	6	88	608	$\mathrm{P_w}$	7	188	2624

illustrated i.e. correct [0,3) (green), poor [3,6) (yellow) and wrong [6, ∞) mm (red). An example of registration with affine and B-spline with 2000 iterations is given in Fig. 5.4a. LSTM 6-3-1 achieved the best performance among the others with only one misclassification out of 5 landmarks in this slice, where it incorrectly predicted

Table 5.4: Detailed hierarchical results of the landmark-based results on the test set, for the trainable LSTM 6-3-1 decoder. We report the aggregated values over all five registration settings: affine and B-spline registration after affine with 20, 100, 500, and 2000 iterations. The sub-indices c, p and w correspond to correct [0,3), poor [3,6) and wrong [6, ∞) mm classes. The shaded cells represent a combination of several fine-grained labels, as in earlier steps more coarse classes are predicted.

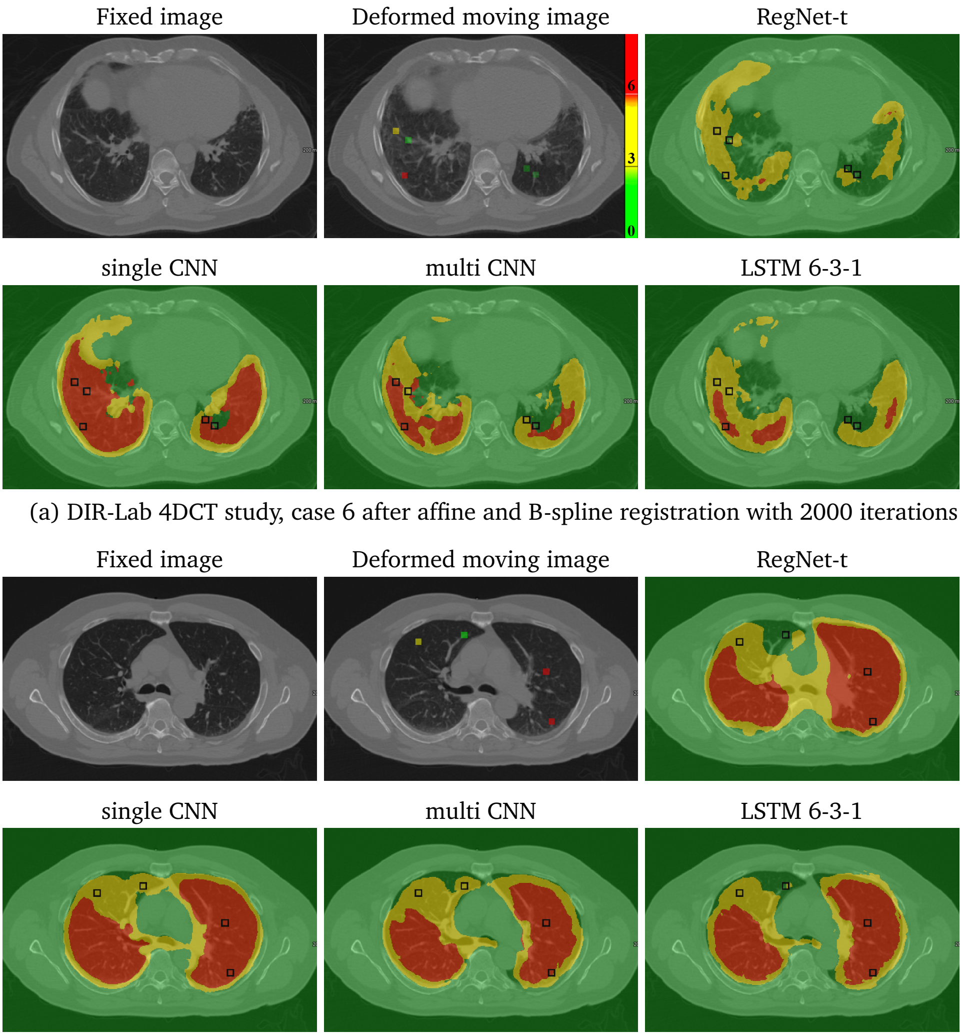
time	$F1_{c\ 0\text{-}1}$	$F1_{c\ 1\text{-}3}$	$F1_p$	$F1_w$	$\overline{F1}$	Acc	κ
SPREAD (case 13 to 21)							
step 1	92.4			60.1	77.1	89.9	0.55
step 2	94.3		53.0	68.9	72.1	83.9	0.66
step 3	23.2	64.6	60.4	87.8	59.0	60.6	0.44
step 3-merged	93.6		60.4	87.8	80.6	87.4	0.75
DIR-Lab 4DCT (case 1 to 10)							
step 1	84.2			14.9	49.6	73.3	0.11
step 2	83.6		28.0	22.9	44.8	61.6	0.32
step 3	53.8	67.2	53.8	73.2	62.0	63.3	0.50
step 3-merged	92.3		53.8	73.2	73.1	79.4	0.66

poor (yellow) label for the correct (green) landmark in the right lung (left side of this image). RegNet-t underpredicted in this slice and misclassified in the wrong (red) regions. Another example with only affine registration is given in Fig. 5.4b. In this slice LSTM 6-3-1 and RegNet-t predicted all four landmarks correctly.

5.3.6 Comparison with Random Forest method

The proposed multi-stage hierarchical LSTM design is compared to a conventional learning-based method using random forests (see Section 5.3.2.5 for details). We compare this method on the SPREAD (cases 13 - 21) and DIR-Lab 4DCT (cases 1 to 5) studies, i.e. we excluded cases 6 to 10 from DIR-Lab 4DCT as these cases were not present in the test set of [33]. Since the random forest method was designed to only predict nonrigid registration error, in this experiment we only included B-spline registrations with 20, 100, 500, and 2000 iterations, thus excluding the affine registration.

The results are reported in Table 5.5. In terms of $\overline{F1}$, the proposed LSTM design achieved significantly better results in both studies. On all F1 measures on both datasets, the LSTM method outperforms the random forest method, except for the $F1_c$ score on the SPREAD study, which were 93.6% vs 96.9% for LSTM vs RF. A compelling advantage of the LSTM method is that it can be applied to affine registrations as well



(a) DIR-Lab 4DCT study, case 6 after affine and B-spline registration with 2000 iterations

(b) DIR-Lab 4DCT study, case 7 after affine registration

Figure 5.4: Examples of the prediction output on entire image pairs registered using conventional registration techniques. The ground truth misalignment on the landmark locations are overlaid in the deformed moving images. These landmarks are dilated in this figure for a better visualization. The color bar indicates the target registration error, which is added on the top center image. For all predictions, a three-label output is illustrated i.e. correct [0,3) (green), poor [3,6) (yellow) and wrong [6, ∞) mm (red). (a) Results on the case 6 from the DIR-Lab 4DCT study. The deformed moving image is obtained after an affine and a B-spline registration with 2000 iterations. (b) Results on the case 7 from the DIR-Lab 4DCT study. The deformed moving image is obtained after an affine transformation.

Table 5.5: Landmark-based results on the overlapping part of the test set, comparing LSTM to the random forests method (RF) [33]. The results include B-spline registration with 20, 100, 500, and 2000 iterations. The sub-indices c, p and w correspond to correct [0,3), poor [3,6) and wrong [6, ∞) mm classes.

method	$F1_c$	$F1_p$	$F1_w$	$\overline{F1}$	Acc
SPREAD (case 13 to 21)					
RF	96.9	40.0	62.4	66.4	92.7
LSTM	93.6	60.4	87.8	80.6	87.4
DIR-Lab 4DCT (case 1 to 5)					
RF	88.2	42.3	34.7	55.1	77.3
LSTM	94.0	56.4	66.7	72.4	84.2

as nonrigid registrations. Another major advantage of the LSTM method is that the inference time is about 22 seconds (for an image size of 410 × 410 × 410 mm) compared to 3 hours for the random forests, where a lot of the time is spent in the feature calculation (registration and local normalized mutual information).

5.4 Discussion

We proposed a deep learning-based method to predict registration misalignment, using a hierarchical LSTM approach with gradual refinements. We performed a wide range of quantitative evaluations on multiple chest CT databases.

The performance of the compared decoders in Table 5.2 are not consistent in all registration settings. The B-spline registration with 2000 iterations represents the most common setting, as this represents an accurate registration. In this case the proposed hierarchical LSTM method achieved the best result in terms of $\overline{F1}$, κ coefficient and the number of cw misclassifications. In the "total" row, the number cw misclassifications of the LSTM method is much smaller than that of the RegNet-t. In the validation set in Table 5.1, the LSTM design achieved slightly better results in comparison to the multi-scale CNN design based on the $\overline{F1}$, κ coefficient and the number of cw misclassifications, showing that utilizing both the multi-resolution approach and hierarchical refinements can improve the misalignment predictions.

The proposed encoding mechanism using RegNet showed to be effective, as it achieved promising results even with a simple thresholding 'decoder' as used in RegNet-t. In predicting the misalignment of the affine registration, RegNet-t outperformed all other decoders. Since RegNet-t resamples images after each stage, potentially it can capture larger registration misalignment. We experimented with a similar setup using the LSTM approach, resampling after each step. However, the results of this experiment were not promising on the validation set. Another difference is that the

Table 5.6: A summary of some of the earlier approaches for estimating registration misalignment. For simplification, results are averaged over all reported test data. RF refers to a random forest and NA refers to “not available”.

article	output	method	data	training	testing	result
Hub et al. [35], 2009	continuous, local	perturbing input	chest CT, in-house	NA	artificial DVF	NA
Muenzing et al. [39], 2012	classification, local	cascade classifiers with intensity based features	chest CT, in-house	landmarks in real data	landmarks in real data	$\overline{\text{F1}}$ 85.2%
Sokooti et al. [33], 2019	regression, local	RF using intensity and registration based features	chest CT, in-house + public	landmarks in real data	landmarks in real data	MAE 1.42 mm, $\overline{\text{F1}}$ 60.75%
Saygili [111], 2020	regression, local	block matching + RF	chest CT, public	landmarks in real data	landmarks in real data	MAE 2.0 mm, Acc 81.8%
Eppenhof et al. [40], 2018	regression, local	CNN	chest CT, public	artificial DVF under 4 mm	landmarks in real data	RMSD 0.66 mm
Senneville et al. [115], 2020	classification, global	CNN + linear regression (classifier)	brain MR, public	artificial affine DVF	real data	Binary Acc 96.0%
Proposed method	classification, local	ConvLSTM	chest CT, in-house + public	artificial DVF under 17 mm	landmarks in real data	$\overline{\text{F1}}$ 76.5%

RegNet was trained on artificial data with a maximum deformation of 20 mm in each direction for the course resolution (RegNet^4), whereas the the maximum deformation in this study is set to 10 mm in each direction (about 17 mm in vector magnitude). It should be noted that in terms of the total number of cw misclassifications, the LSTM and CNN designs are still more in favor, which are reported as 9, 2, and 12 for the LSTM, multi-scale CNN and RegNet-t, in order (see the first four rows in Table 5.2).

The distribution of the labels “correct”, “poor” and “wrong” are highly imbalanced in image registration. For instance, in the test set within five registration settings, the distribution of samples are 67.8%, 14.7%, 17.5% in the SPREAD study and 53.5%, 17.5%, 29.0% in the DIR-Lab 4DCT for the labels correct, poor and wrong, respectively. In order to mimic the same distribution during training, the probability of selecting patches in the range [0,3), [3,6) and $[6, \infty)$ mm are set to 60%, 20% and 20%, respectively (see Section 5.3.2.1). However, this can influence the first step of the LSTM training as the sampling becomes imbalanced again in this step.

A comparison to previous methods for predicting registration misalignment is not trivial due to differences in approach (classification, regression) as well as the use of different test datasets. Table 5.6 gives an overview of several methods from the literature. A classification-based approach to estimate registration misalignment was also presented in [39]. They proposed a classical learning-based approach using several hand-crafted features. Muenzing et al. [39] reported F1 scores of 95.3%, 73.8% and 86.6% in the labels [0,2), [2,5) and $[5, \infty)$ mm. It is not trivial to compare our results to this method because the evaluation is done on different data and using different thresholds for labels. When it comes to the dense prediction for an entire image, calculating those hand-crafted features become quite time-consuming. In the CNN-based approaches, Eppenhof et al. [40] proposed a regression network to predict registration misalignment. They trained on the odd-numbered images from the DIR-Lab-4DCT and the COPDgene data sets and tested on the even-numbered scans, and on two additional chest CT studies. They reported a root-mean-square deviation (RMSD) of 0.66 mm between the ground truth TRE and the predicted one for landmarks with ground truth TRE below 4 mm. The main limitation is that the method predicts registration misalignment smaller than 4 mm only. Since our proposed method has one label corresponding to misalignment in the range $[6, \infty)$ mm, a quantitative comparison is not feasible. In Section 5.3.5, we drew a comparison between the proposed LSTM method and a random forests regression method [33]. We kept the experiment settings as similar as possible. However, some minor differences still exist. For instance, the voxel size in the LSTM method is resampled to an isotropic size of [1, 1, 1] mm, whereas in the random forests method, resampling is not applied. Since one of the proposed features in [33] was the variation of the transformations with respect to the initial states of the B-spline grid, it is not possible to use this approach for affine registration.

In this study, we proposed to use RegNet [18] to encode a pair of images using a multi-resolution approach to high-dimensional feature maps. Although the experiment with a simple decoder as RegNet-t reveals that encoding with RegNet is quite powerful, potentially, any registration network can be used instead of RegNet. It could therefore be interesting to perform a comparison between different network architectures. The proposed method is designed with three resolutions of the input given in three steps to the LSTM block. At the third resolution, the receptive field of the network is usually larger than an entire chest CT image (with a spacing of 1 mm). Thus, potentially no further contextual information can be achieved by increasing the number of resolutions. However, varying the number of steps in the LSTM block can be an interesting experiment. We experimented with three steps, but with various splitting values in Section 5.3.4. The number of steps of the LSTM can be increased even with identical inputs, similar to [121].

The proposed method is expected to be sensitive to anatomical changes like tumor growth. Thus, it may detect those regions as a suboptimal local registration. This limitation may potentially be addressed by adding a new type of deformation to the artificial training data strategy, which mimics such anatomical changes. For example, in this study we modelled respiratory motion specifically designed for lungs (see Section 5.2.2), as we performed all experiments on chest CT scans. This may be extended with additional realistic artificial data generation types, for other use cases.

However, the proposed training and prediction methods are generic and independent of the image type. In future work, the proposed method could be evaluated on other modalities and anatomical sites as well. Although all nonrigid experiments in this study are performed using B-spline registration, potentially, the proposed method is independent of the registration paradigm and can be applied to other nonrigid registration methods.

5.5 Conclusion

We proposed a framework for classifying registration misalignment using deep learning, consisting of encoding relevant features in a latent space and a hierarchical and gradually refining LSTM decoder for the prediction. Multi-resolution contextual information is incorporated in the design. The network is fully trained over artificially generated images, while the evaluation is performed over realistic chest CT scans. The proposed decoder is compared with two other CNN-based decoders and a method based on the output of a deep learning based registration RegNet-t. A comprehensive study is performed on two independent test sets (SPREAD case 13 to 21, and DIR-Lab 4DCT) with various registration settings. In the B-spline registration with 2000 iterations, the proposed method achieved an $\overline{\mathrm{F1}}$ and number of cw misclassifications of 68.9%, 2 and 63.9%, 22 in the SPREAD and the DIR-LAB 4DCT studies, respectively. In the aggregation of all registration settings, the proposed LSTM design obtained the least number of cw misclassifications. At the inference time, the proposed method can predict a dense map in about 22 seconds.

6

Summary and Future Work

Image registration is a crucial task in medical image processing. Performing an automatic fast image registration with less manual finetuning can speed up numerous medical image processing procedures. In addition, an automatic quality assessment of registration can speed up this time-consuming task. In this thesis, we developed a learning-based image registration technique called RegNet. Moreover, we proposed two quality assessment mechanisms using random forests (RF) and convolutional long short term memory (ConvLSTM), in which the latter performs faster and more accurate. In this chapter, we summarize the previous chapters and discuss potential directions of future research.

6.1 Summary

In the first chapter, we provided general information about image registration and quality assessment of registration. In Chapter 2, we propose a convolutional neural network architecture to solve nonrigid image registration through a learning approach. The proposed RegNet is trained using a set of random artificially generated DVFs with a maximum deformation of 8 mm in each direction. In Chapter 3, we substantially improve the proposed RegNet by utilizing a multi-stage approach and improving the artificial data generation procedure. A quantitative error prediction of medical image registration is proposed in Chapter 4 using regression forests. The forest is built with features related to the transformation model and features related to the dissimilarity after registration on distinctive landmark locations. In Chapter 5, a hierarchical prediction of registration misalignment using a convolutional LSTM with application to chest CT scans is proposed. The proposed method is substantially faster than methods involving multiple registrations.

Chapter 2 In this chapter, we propose a method to solve nonrigid image registration through a learning approach, instead of via iterative optimization of a predefined dissimilarity metric. We design a Convolutional Neural Network (CNN) architecture that, in contrast to all other work, directly estimates the displacement vector field (DVF) from a pair of input images. This chapter is one of the first proposed methods in nonrigid DL-based image registrations. The proposed RegNet is trained using a set of random artificially generated DVFs with a maximum deformation of 8 mm in each direction.The proposed method does not explicitly define a dissimilarity metric, and integrates image content at multiple scales to equip the network with contextual information. At testing time nonrigid registration is performed in a single shot, in contrast to current iterative methods. We tested RegNet on an in-house chest CT study called SPREAD. It achieved an average target registration error (TRE) of 1.66 mm over the test cases. The results show that the accuracy of RegNet is on par with a conventional B-spline registration, for anatomy within the capture range, i.e. less than 8 mm.

In **Chapter 3**, we substantially enhance the initial RegNet method introduced in Chapter 2. The newly proposed method utilizes a multi-stage approach, which significantly enlarges the capture range. The artificial data generation was improved by including more generic deformations as well as more realistic deformations like respiratory motion. We experimented with various network architectures and the proposed "U-Net-advanced" design achieved better performance in the validation set. This design was similar to a U-Net, but with addition of dilated convolutional layers. The proposed method, RegNet, is evaluated on multiple databases of chest CT scans and achieved a target registration error of 2.32 ± 5.33 mm and 1.86 ± 2.12 mm on SPREAD and DIR-Lab-4DCT studies, respectively. Consequently, the enhanced RegNet achieved the best result on the DIR-Lab 4DCT study among all published DL-based registration methods. The average inference time of RegNet with two stages is about 2.2 s.

Chapter 4 presents a quantitative error prediction of medical image registration using regression forests. A new automatic method is proposed to predict the registration error in a quantitative manner, and is applied to chest CT scans. A random regression forest is utilized to predict the registration error locally. The forest is built with features related to the transformation model and features related to the dissimilarity after registration. The feature set consists of the variation of displacement vector field, the coefficient of variation of joint histograms, determinant of the Jacobian, the modality independent neighborhood descriptor (MIND), and the local normalized mutual information. The forest is trained and tested using manually annotated corresponding points between pairs of chest CT scans in two experiments: SPREAD (trained and tested on SPREAD) and inter-database (including three databases SPREAD, DIR-Lab-4DCT and DIR-Lab-COPDgene). The results show that the mean absolute errors of regression are 1.07 ± 1.86 and 1.76 ± 2.59 mm for the SPREAD and inter-database experiment, respectively. The overall accuracy of classification in three classes (correct, poor, and wrong registration) is 90.7% and 75.4%, for SPREAD and inter-database respectively. The good performance of the proposed method enables important applications such as automatic quality control in large-scale image analysis.

In **Chapter 5**, a hierarchical prediction of registration misalignment using a convolutional LSTM with application to chest CT scans is proposed. The proposed method is substantially faster than methods involving multiple registrations. This task is casted to a classification problem with multiple classes of misalignment: "correct" 0-3 mm, "poor" 3-6 mm and "wrong" over 6 mm. Rather than a direct prediction, we propose a hierarchical approach, where the prediction is gradually refined from coarse to fine. Our solution is based on a convolutional Long Short Term Memory (LSTM), using hierarchical misalignment predictions on three resolutions of the image pair, leveraging the intrinsic strengths of an LSTM for this problem. The convolutional LSTM is trained on a set of artificially generated image pairs obtained from artificial displacement vector fields (DVFs). Results on chest CT scans show that incorporating multi-resolution information, and the hierarchical use via an LSTM for this, leads to overall better F1 scores, with fewer misclassifications in a well-tuned registration setup. The final system yields an accuracy of 87.1%, and an average F1 score of 66.4% aggregated in two independent chest CT scan studies.

6.2 Discussion and Future Work

The work presented in this thesis was aimed at developing methods to perform image registration as well as quality assessment of image registrations.

In the proposed RegNet in Chapters 2 and 3, we utilized a deep convolutional neural network approach. Although deep learning methods in segmentation applications achieved promising results, several challenges still exist in the registration applications. Finding the optimal solution in conventional segmentation methods like level set and min cost (minimum of the cost function) are usually iterative similar to the conventional image registration techniques. However, in conventional image segmentations, the main image is constant in all iterations and the segmentation is updated in each iteration. On the contrary, in the conventional iterative image registration, an implicit (or explicit) resampling of the deformed moving image is performed at each iteration. Apparently, predicting the final transformation in one shot is still challenging in DL-based methods. We noticed significant improvement when using the multi-stage approach, in which a resampling was also performed. As reported in Table 2.2 the average TRE was improved from 3.80 mm to 1.57 mm. It is worth noting that the registration quality may still be improved by sequentially employing multiple RegNets in the original resolution. Potentially, a simple stopping criterion like the difference of variation between subsequent transformations or a more complex approach such as reinforcement learning can be used.

We utilized a supervised transformation approach in Chapter 2 and 3 of this thesis. Quite a few articles proposed an unsupervised transformation approach [23, 24, 8]. One of the advantages of the unsupervised transformation method is that the training data can be fully realistic including fully realistic ground truth transformations. Usually the training is performed with a simple dissimilarity metric like mutual information. Thus, a potential disadvantage of unsupervised methods is that the ground truth transformation is not known, and the trained network may not necessarily be better than a conventional iterative registration with the same dissimilarity metric. On the other hand, in the supervised transformation approach, the ground truth transformation is accurate (not necessarily unique as registration is usually an ill-posed problem). On the contrary, the transformation and usually one of the images (the deformed moving image) may not be completely realistic. Although artificial data generation could be a potential way to get higher performance than human experts, this could be too idealistic at this moment, where current deep learning based registrations could be much further enhanced. All in all, in order to improve the supervised approaches, the necessity for a large medical dataset providing ground truth for transformations can be strongly felt. This can be done by annotating distinctive landmarks as well as region segmentations.

In Chapter 4, we proposed a regression approach to predict the registration error and in Chapter 5, we simplified the task into a classification approach. Each of these approaches have their owns pros and cons. It should be highlighted that in a regression approach, the acceptable margin of the regression error correlates with the ground truth value. Thus, a normalized loss could help the training to converge better. In the classification approach, this issue is eliminated when considering a class with large values like $[6,\infty)$ mm. However, it should be pointed out that when defining the labels as correct, $[0,3)$ mm, poor $[3,6)$ mm, and wrong $[6,\infty)$ mm, the labels are not ordinal anymore but more similar to ordered labels. This means that a mis-classification between the correct and wrong label is worse than a mis-classification between two adjacent labels such as correct and poor. This criterion does not naturally exist in classification approaches but can be imposed using hierarchical classification. In general, the classification is more difficult for values close to the border. For instance, it is not trivial to classify a value of 2.99 in either the correct or the poor class. A solution might be to utilize soft ground truth labels, for example for the value 2.99, the ground truth can be set to [0.60, 0.40, 0] for classes correct, poor, and wrong respectively. In the hyperspherical prototype approach [131] the one hot encoded ground truth will be mapped to an output space. This way we can provide *a priori* information about the labels and utilize that in the organization of the output space. For instance, the wrong and the poor labels can be close to each other, while the wrong and the correct labels can have more distance. Another interesting application in the hyperspherical prototype is to simultaneously perform the regression and classification even in the same output space.

In Chapters 2, 3, and 5, we utilize artificial data generations in order to train a convolutional neural network to learn registration or registration error. We proposed “single frequency”, “mixed frequency”, and “respiratory motion” approaches to artificially generate displacement vector fields. One of the limitations of the aforementioned generations is that they are expected to be sensitive to anatomical changes like tumor growth. This limitation may potentially be addressed by adding a new type of deformation to the artificial training data strategy, which mimics such anatomical changes. In general, the artificial generation can be further enhanced by adding more realistic and complex simulations. For instance, if a rib segmentation is available in chest CT scans, it is possible to perform nonrigid deformation outside of the rib and rigid deformations inside the rib. The network potentially can learn the relation between organs and the rigidity of the deformations. Other realistic deformations like sliding motion of the lungs can also be added to the training images.

Although all experiments in this thesis are performed in chest CT scans, all proposed methods are generic and potentially can be applied to other modalities and anatomical sites as well. In a similar study on intrasubject magnetic resonance brain images registration, RegNet was trained on brain MR images and showed promising results [21]. However, utilizing artificial data generation in multi-modality images need to be investigated in the future, as potentially an intensity mapping approach [132] might be needed.

6.3 General conclusions

In conclusion, this thesis proposes learning-based methods for medical image registration and for quality assessment of image registration. All proposed methods are fully automatic and do not require human interactions. The proposed RegNet architecture was tested on registration chest CT scan pairs and achieved on par results with a conventional B-spline registration method. The hierarchical classification framework to detect registration misalignment using long short term memory convolutional neural networks (ConvLSTM) obtained promising results. All deep learning methods described in this thesis have a runtime in the order of seconds, substantially improving over conventional methods.

Samenvatting en toekomstig werk

Beeldregistratie is een cruciale taak in de medische beeldverwerking. Het uitvoeren van een snelle automatische beeldregistratie met minder handmatige fijnafstelling kan tal van medische beeldverwerkingsprocedures versnellen. Bovendien kan een automatische kwaliteitsbeoordeling van de registratie deze tijdrovende taak versnellen. In dit proefschrift hebben we een op leren gebaseerde beeldregistratietechniek ontwikkeld, genaamd RegNet. Bovendien hebben we twee kwaliteitsbeoordelingsmechanismen voorgesteld, gebruikmakend van random forests (RF) en convolutional long short term memory (ConvLSTM), waarbij de laatste sneller en nauwkeuriger presteert. In dit hoofdstuk vatten we de vorige hoofdstukken samen en bespreken we mogelijke richtingen voor vervolgonderzoek.

Samenvatting

In het eerste hoofdstuk hebben we algemene informatie gegeven over beeldregistratie en kwaliteitsbeoordeling van registratie. In Hoofdstuk 2 stellen we een convolutional neural network architectuur voor om niet-rigide beeldregistratie op te lossen door middel van een leerbenadering. Het voorgestelde RegNet wordt getraind met behulp van een reeks willekeurig kunstmatig gegenereerde DVF's met een maximale vervorming van 8 mm in elke richting. In Hoofdstuk 3 hebben we het voorgestelde RegNet substantieel verbeterd door een meerfasige benadering te gebruiken en de kunstmatige procedure voor het genereren van de DVF's te verbeteren. Een kwantitatieve foutvoorspelling van medische beeldregistratie wordt voorgesteld in Hoofdstuk 4 met behulp van een regressie forest. Deze random forest is opgebouwd met kenmerken die betrekking hebben op het transformatiemodel en kenmerken die verband houden met de ongelijkheid na registratie op onderscheidende landmark locaties. In Hoofdstuk 5 wordt een hiërarchische voorspelling van de registratiefout voorgesteld met behulp van een convolutional LSTM met toepassing op CT scans van de borstkas. De voorgestelde methode is aanzienlijk sneller dan methoden die meerdere registraties nodig hebben.

Hoofdstuk 2 In dit hoofdstuk stellen we een methode voor om niet-rigide beeldregistratie op te lossen door middel van een leerbenadering, in plaats van via iteratieve optimalisatie van een vooraf gedefinieerde ongelijkheidsmetriek. We ontwerpen een Convolutional Neural Network (CNN)-architectuur die, in tegenstelling tot al het andere werk, het verplaatsingsvectorveld(DVF) rechtstreeks schat op basis van een paar invoerbeelden. Dit hoofdstuk is een van de eerste voorgestelde methoden voor niet-rigide DL-gebaseerde beeldregistratie. Het voorgestelde RegNet wordt getraind met behulp van een reeks willekeurig kunstmatig gegenereerde DVF's met een maximale vervorming van 8 mm in elke richting. De voorgestelde methode definieert niet expliciet een ongelijkheidsmetriek en integreert beeldinhoud op meerdere schalen om het network uit te rusten met contextuele informatie. Tijdens het testen wordt niet-rigide registratie in één keer uitgevoerd, in tegenstelling tot de huidige iteratieve methoden. We hebben RegNet getest met een interne dataset van CT scans van de borstkas, SPREAD genaamd. RegNet behaalde een gemiddelde doelregistratiefout (Target Registration Error, TRE) van 1,66 mm over de testcases. De resultaten laten zien dat de nauwkeurigheid van RegNet vergelijkbaar is met een conventionele B-spline-registratie, voor anatomie binnen het bereik, d.w.z. minder dan 8 mm.

In **Hoofdstuk 3** verbeteren we de oorspronkelijke RegNet-methode die in Hoofdstuk 2 is geïntroduceerd aanzienlijk. De nieuw voorgestelde methode maakt gebruik van een meertrapsbenadering, waardoor het bereik aanzienlijk wordt vergroot. De kunstmatige generatie van de DVF's is verbeterd door meer algemene vervormingen op te nemen, evenals meer realistische vervormingen zoals ademhalingsbewegingen. We hebben geëxperimenteerd met verschillende network architecturen en het voorgestelde "U-Net-advanced"-ontwerp leverde betere prestaties in de validatieset. Dit ontwerp was vergelijkbaar met een U-Net, maar met toevoeging van gedilateerde convolutional lagen. De voorgestelde methode, RegNet, wordt geëvalueerd op meerdere databases van CT-scans van de borstkas en behaalde een doelregistratiefout van $2,32 \pm 5,33$ mm en $1,86 \pm 2,12$ mm op SPREAD en DIR-Lab- 4DCT-onderzoeken, respectievelijk. Bijgevolg behaalde het verbeterde RegNet het beste resultaat in de DIR-Lab 4DCT-studie van alle gepubliceerde op DL-gebaseerde registratiemethoden. De gemiddelde inferentietijd van RegNet met twee fasen is ongeveer 2,2 s.

Hoofdstuk 4 presenteert een kwantitatieve foutvoorspelling van medische beeldregistratie met behulp van een regressie forest. Een nieuwe automatische methode wordt voorgesteld om de registratiefout op een kwantitatieve manier te voorspellen, en wordt toegepast op CT-scans van de borstkas. Om de registratiefout lokaal te voorspellen, wordt gebruik gemaakt van een random regressie forest. Het forest is gebouwd met kenmerken die betrekking hebben op het transformatiemodel en kenmerken die verband houden met de ongelijkheid na registratie. De kenmerkenset bestaat uit de variatie van het verplaatsingsvectorveld, de variatiecoëfficiënt van gezamenlijke histogrammen, determinant van de Jacobiaan, de modality independent neighborhood descriptor (MIND) en de lokale genormaliseerde wederzijdse informatie. Het forest wordt getraind en getest met behulp van handmatig geannoteerde corresponderende punten tussen paren CT-scans van de borstkas in twee experimenten: SPREAD (getraind en getest op SPREAD) en inter-database (met drie databases SPREAD, DIR-Lab-4DCT en DIR-Lab-COPDgene). De resultaten laten zien dat de gemiddelde absolute regressiefouten 1,07 ± 1,86 en 1,76 ± 2,59 mm zijn voor respectievelijk het SPREAD- en het interdatabase-experiment. De algehele nauwkeurigheid van classificatie in drie klassen (correct, slecht en verkeerde registratie) is 90,7% en 75,4%, voor respectievelijk SPREAD en inter-database. De goede prestaties van de voorgestelde methode maken belangrijke toepassingen mogelijk, zoals automatische kwaliteitscontrole bij grootschalige beeldanalyse.

In **Hoofdstuk 5** wordt een hiërarchische voorspelling van verkeerde uitlijning van de registratie voorgesteld met behulp van een convolutional LSTM met toepassing op CT-scans van de borstkas. De voorgestelde methode is aanzienlijk sneller dan methodes met meerdere registraties. Deze taak wordt geprojecteerd naar een classificatieprobleem met meerdere klassen van verkeerde uitlijning: “correct” 0-3 mm, “slecht” 3-6 mm en “verkeerd” meer dan 6 mm. In plaats van een directe voorspelling stellen we een hiërarchische benadering voor, waarbij de voorspelling geleidelijk wordt verfijnd van grof naar fijn. Onze oplossing is gebaseerd op een convolutional Long Short Term Memory (LSTM), dat hiërarchische voorspellingen van verkeerde uitlijning gebruikt op drie resoluties van het beeldpaar, waarbij gebruik wordt gemaakt van de intrinsieke sterke punten van een LSTM voor dit probleem. De convolutional LSTM wordt getraind op een set kunstmatig gegenereerde beeldparen die zijn verkregen uit kunstmatige verplaatsingsvectorvelden (DVF’s). Resultaten op CT-scans van de borstkas laten zien dat het opnemen van informatie met meerdere resoluties, en het hiërarchische gebruik via een LSTM hiervoor, leidt tot over het algemeen betere F1-scores, met minder misclassificaties in een goed afgestemde registratieconfiguratie. Het uiteindelijke systeem levert een nauwkeurigheid op van 87, 1% en een gemiddelde F1-score van 66, 4%, gemiddeld over twee onafhankelijke CT-scanonderzoeken van de borstkas.

Discussie en vervolgonderzoek

Het werk dat in dit proefschrift wordt gepresenteerd, was gericht op het ontwikkelen van methoden voor het uitvoeren van beeldregistratie en het beoordelen van de kwaliteit van beeldregistraties.

In het voorgestelde RegNet in de hoofdstukken 2 en 3 hebben we een diepe convolutional neural network benadering gebruikt. Hoewel deep learning-methoden in segmentatietoepassingen veelbelovende resultaten hebben opgeleverd, bestaan er nog steeds verschillende uitdagingen in de registratietoepassingen. Het vinden van de optimale oplossing in conventionele segmentatiemethoden zoals level set en min cost (minimum van de kostenfunctie) is meestal iteratief, vergelijkbaar met conventionele beeldregistratietechnieken. Bij conventionele beeldsegmentaties is het hoofdbeeld echter constant in alle iteraties en wordt de segmentatie in elke iteratie bijgewerkt. In tegendeel, bij de conventionele iteratieve beeldregistratie wordt bij elke iteratie een impliciete (of expliciete) herbemonstering van het vervormde bewegende beeld uitgevoerd. Blijkbaar is het voorspellen van de uiteindelijke transformatie in één keer nog steeds een uitdaging in op DL gebaseerde methoden. We merkten significante verbetering bij het gebruik van de meertrapsbenadering, waarbij ook een resampling werd uitgevoerd. Zoals vermeld in Tabel 2.2 is de gemiddelde TRE verbeterd van 3,80 mm naar 1,57 mm. Het is vermeldenswaard dat de registratiekwaliteit nog kan worden verbeterd door achtereenvolgens meerdere RegNets in de oorspronkelijke resolutie te gebruiken. Mogelijk kan een eenvoudig stopcriterium worden gebruikt, zoals het verschil in variatie tussen opeenvolgende transformaties of een meer complexe benadering, zoals reinforcement learning.

In Hoofdstuk 2 en 3 van dit proefschrift hebben we gebruik gemaakt van een gesuperviseerde transformatiebenadering. Heel wat artikelen stelden een niet-gesuperviseerde transformatiebenadering voor [23, 24, 8]. Een van de voordelen van de niet-gesuperviseerde transformatiemethode is dat de trainingsgegevens volledig realistisch kunnen zijn, inclusief volledig realistische ground truth transformaties. Gewoonlijk wordt de training uitgevoerd met een eenvoudige ongelijkheidsmetriek zoals mutual information. Een mogelijk nadeel van methoden zonder supervisie is dus dat de transformatie van de ground truth niet bekend is en dat het getrainde network niet noodzakelijk beter is dan een conventionele iteratieve registratie met dezelfde metriek voor ongelijkheid. Aan de andere kant, in de gesuperviseerde transformatiebenadering, is de ground truth transformatie accuraat (niet noodzakelijk uniek aangezien registratie meestal een slecht gesteld probleem is). Integendeel, de transformatie en meestal een van de beelden (het vervormde bewegende beeld) is misschien niet helemaal realistisch. Hoewel kunstmatige gegevensgeneratie een potentiële manier zou kunnen zijn om betere prestaties te krijgen dan menselijke experts, zou dit op dit moment te idealistisch kunnen zijn, waar de huidige op deep learning gebaseerde registraties veel verder zouden kunnen worden verbeterd. Al met al, om de gesuperviseerde benaderingen te verbeteren, kan de noodzaak van een grote medische dataset met een ground truth voor transformaties sterk worden gevoeld. Dit kan worden gedaan door onderscheidende oriëntatiepunten of regiosegmentaties te annoteren.

In Hoofdstuk 4 hebben we een regressiebenadering voorgesteld om de registratiefout te voorspellen en in Hoofdstuk 5 hebben we de taak vereenvoudigd tot een classificatiebenadering. Elk van deze benaderingen heeft zijn eigen voor- en nadelen. Benadrukt moet worden dat bij een regressiebenadering de aanvaardbare marge van de regressiefout correleert met de ground truth waarde. Een genormaliseerd verlies zou dus kunnen helpen om de training beter te laten convergeren. In de classificatiebenadering wordt dit probleem geëlimineerd door het beschouwen van een klasse met grote waarden zoals $[6,\infty)$ mm. Er moet echter op worden gewezen dat bij het definiëren van de labels als correct, $[0,3)$ mm, slecht $[3,6)$ mm, en verkeerd $[6,\infty)$ mm, de labels niet meer ordinaal zijn maar meer lijken op geordende labels. Dit betekent dat een verkeerde classificatie tussen het correcte en een verkeerde label erger is dan een verkeerde classificatie tussen twee aangrenzende labels, zoals correct en slecht. Dit criterium komt van nature niet voor in classificatiebenaderingen, maar kan worden opgelegd met behulp van hiërarchische classificatie. Over het algemeen is de classificatie moeilijker voor waarden dicht bij de grens. Het is bijvoorbeeld niet triviaal om een waarde van 2,99 in de correcte of de slechte klasse in te delen. Een oplossing zou kunnen zijn om zachte ground truth labels te gebruiken, bijvoorbeeld voor de waarde 2,99 kan de ground truth worden ingesteld op [0,60, 0,40, 0] voor respectievelijk de klassen correct, slecht en verkeerd. In de hypersferische prototypebenadering [131] zal de one-hot gecodeerde ground truth worden toegewezen aan een outputruimte. Op deze manier kunnen we *a priori* informatie geven over de labels en die gebruiken in de organisatie van de outputruimte. Zo kunnen de verkeerde en de slechte labels dicht bij elkaar liggen, terwijl de verkeerde en de correcte labels meer afstand kunnen hebben. Een andere interessante toepassing in het hypersferische prototype is om tegelijkertijd de regressie en de classificatie uit te voeren, zelfs in dezelfde outputruimte.

In Hoofdstukken 2, 3 en 5 gebruiken we kunstmatige datageneratie om een convolutional neural network te trainen om registratie of registratiefouten te leren. We stelden benaderingen voor “single frequency”, “mixed frequency” en “respiratory motion” voor om kunstmatig verplaatsingsvectorvelden te genereren. Een van de beperkingen van de bovengenoemde generaties is dat ze naar verwachting gevoelig zijn voor anatomische veranderingen zoals tumorgroei. Deze beperking kan mogelijk worden aangepakt door een nieuw type vervorming toe te voegen aan de kunstmatige trainingsgegevensstrategie, die dergelijke anatomische veranderingen nabootst. Over het algemeen kan de kunstmatige generatie verder worden verbeterd door meer realistische en complexe simulaties toe te voegen. Als er bijvoorbeeld een ribsegmentatie beschikbaar is in CT-scans van de borstkas, is het mogelijk om niet-rigide vervorming buiten de rib en rigide vervormingen binnen de rib uit te voeren. Het network kan potentieel de relatie tussen organen en de stijfheid van de vervormingen leren. Andere realistische vervormingen zoals glijdende beweging van de longen kunnen ook aan de trainingsbeelden worden toegevoegd.

Hoewel alle experimenten in dit proefschrift worden uitgevoerd met CT-scans van de borstkas, zijn alle voorgestelde methoden generiek en kunnen ze mogelijk ook worden toegepast op andere modaliteiten en anatomische locaties. In een soortgelijk onderzoek naar de registratie van intra-subject MR hersenbeelden, werd RegNet getraind op MR-beelden van de hersenen en liet het veelbelovende resultaten zien [21]. Het gebruik van kunstmatige datageneratie in afbeeldingen met meerdere modaliteiten moet in de toekomst echter worden onderzocht, omdat mogelijk een benadering voor het tranformeren van de intensiteit [132] nodig kan zijn.

Algemene conclusies

Concluderend stelt dit proefschrift leergebaseerde methoden voor ten behoeve van medische beeldregistratie en kwaliteitsbeoordeling van beeldregistratie. Alle voorgestelde methoden zijn volledig automatisch en vereisen geen menselijke interactie. De voorgestelde RegNet-architectuur werd getest op CT-scanparen van de borstkas en behaalde resultaten vergelijkbaar met een conventionele B-spline registratiemethode. Het hiërarchische classificatieraamwerk om onjuiste uitlijning van registratie te detecteren met behulp van long short term memory convolutional neural networks (ConvLSTM) heeft veelbelovende resultaten opgeleverd. Alle deep learning-methoden die in dit proefschrift worden beschreven, hebben een uitvoeringstijd in de orde van seconden, wat aanzienlijk beter is dan conventionele methoden.

Bibliography

[1] F. P. Oliveira and J. M. R. Tavares. “Medical image registration: a review”. In: *Computer methods in biomechanics and biomedical engineering* 17.2 (2014), pages 73–93.

[2] V. Fortunati, R. F. Verhaart, F. Angeloni, A. Van Der Lugt, W. J. Niessen, J. F. Veenland, M. M. Paulides, and T. Van Walsum. “Feasibility of multimodal deformable registration for head and neck tumor treatment planning”. In: *International Journal of Radiation Oncology* Biology* Physics* 90.1 (2014), pages 85–93.

[3] H. Sokooti. “Fluorescein Angiography in the Diagnosis of Diabetic Retinopathy”. Master’s thesis. K. N. Toosi University of Technology, 2014.

[4] Z. Sun, Y. Qiao, B. P. Lelieveldt, M. Staring, A. D. N. Initiative, et al. “Integrating spatial-anatomical regularization and structure sparsity into SVM: Improving interpretation of Alzheimer’s disease classification”. In: *NeuroImage* 178 (2018), pages 445–460.

[5] F. L. Bookstein. “Principal warps: Thin-plate splines and the decomposition of deformations”. In: *IEEE Transactions on pattern analysis and machine intelligence* 11.6 (1989), pages 567–585.

[6] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. Hill, M. O. Leach, and D. J. Hawkes. “Nonrigid registration using free-form deformations: application to breast MR images”. In: *IEEE Transactions on Medical Imaging* 18.8 (1999), pages 712–721.

[7] S. Ruder. “An overview of gradient descent optimization algorithms”. In: *arXiv preprint arXiv:1609.04747* (2016).

[8] M. S. Elmahdy, J. M. Wolterink, H. Sokooti, I. Išgum, and M. Staring. “Adversarial Optimization for Joint Registration and Segmentation in Prostate CT Radiotherapy”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. Springer International Publishing, 2019, pages 366–374.

[9] W. Grimson, G. Ettinger, S. White, T. Lozano-Perez, W. Wells, and R. Kikinis. “An automatic registration method for frameless stereotaxy, image guided surgery, and enhanced reality visualization”. In: *IEEE Transactions on Medical Imaging* 15.2 (1996), pages 129–140.

[10] G. Haskins, U. Kruger, and P. Yan. “Deep learning in medical image registration: a survey”. In: *Machine Vision and Applications* 31.1 (2020), pages 1–18.

[11] A. Sedghi, J. Luo, A. Mehrtash, S. Pieper, C. M. Tempany, T. Kapur, P. Mousavi, and W. M. W. III. “Semi-supervised image registration using deep learning”. In: *Medical Imaging 2019: Image-Guided Procedures, Robotic Interventions, and Modeling*. Volume 10951. International Society for Optics and Photonics. SPIE, 2019, pages 371 –376.

[12] G. Haskins, J. Kruecker, U. Kruger, S. Xu, P. A. Pinto, B. J. Wood, and P. Yan. "Learning deep similarity metric for 3D MR–TRUS image registration". In: *International journal of computer assisted radiology and surgery* 14.3 (2019), pages 417–425.

[13] D. Mattes, D. R. Haynor, H. Vesselle, T. K. Lewellen, and W. Eubank. "PET-CT image registration in the chest using free-form deformations". In: *IEEE Transactions on Medical Imaging* 22.1 (2003), pages 120–128.

[14] M. P. Heinrich, M. Jenkinson, M. Bhushan, T. Matin, F. V. Gleeson, S. M. Brady, and J. A. Schnabel. "MIND: Modality independent neighbourhood descriptor for multi-modal deformable registration". In: *Medical Image Analysis* 16.7 (2012). Special Issue on the 2011 Conference on Medical Image Computing and Computer Assisted Intervention, pages 1423–1435.

[15] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[16] R. Liao, S. Miao, P. de Tournemire, S. Grbic, A. Kamen, T. Mansi, and D. Comaniciu. "An Artificial Agent for Robust Image Registration". In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. AAAI'17. San Francisco, California, USA: AAAI Press, 2017, 4168–4175.

[17] H. Sokooti, B. De Vos, F. Berendsen, B. P. Lelieveldt, I. Išgum, and M. Staring. "Nonrigid image registration using multi-scale 3D convolutional neural networks". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Volume 10433. Lecture Notes in Computer Science. 2017, pages 232–239.

[18] H. Sokooti, B. de Vos, F. Berendsen, M. Ghafoorian, S. Yousefi, B. P. Lelieveldt, I. Isgum, and M. Staring. "3D convolutional neural networks image registration based on efficient supervised learning from artificial deformations". In: *arXiv preprint arXiv:1908.10235* (2019).

[19] K. A. Eppenhof, M. W. Lafarge, P. Moeskops, M. Veta, and J. P. Pluim. "Deformable image registration using convolutional neural networks". In: *Medical Imaging 2018: Image Processing*. Volume 10574. International Society for Optics and Photonics. 2018, 105740S.

[20] J. Fan, X. Cao, P.-T. Yap, and D. Shen. "BIRNet: Brain image registration using dual-supervised fully convolutional networks". In: *Medical Image Analysis* 54 (2019), pages 193–206.

[21] C. Ji. "Nonrigid image registration using 3D convolutional neural network with application to brain MR images". Master's thesis. Delft, the Netherlands: Delft University of Technology, 2019.

[22] B. D. de Vos, F. F. Berendsen, M. A. Viergever, M. Staring, and I. Išgum. "End-to-end unsupervised deformable image registration with a convolutional neural network". In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2017, pages 204–212.

[23] B. de Vos, F. Berendsen, M. A. Viergever, H. Sokooti, M. Staring, and I. Išgum. “A deep learning framework for unsupervised affine and deformable image registration”. In: *Medical Image Analysis* 52 (2019), pages 128–143.

[24] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca. “An unsupervised learning model for deformable medical image registration”. In: *arXiv preprint arXiv:1802.02604* (2018).

[25] D. Mahapatra, Z. Ge, S. Sedai, and R. Chakravorty. “Joint registration and segmentation of Xray images using generative adversarial networks”. In: *International Workshop on Machine Learning in Medical Imaging*. Springer. 2018, pages 73–80.

[26] A. Hering, S. Häger, J. Moltz, N. Lessmann, S. Heldmann, and B. van Ginneken. “CNN-based lung CT registration with multiple anatomical constraints”. In: *Medical Image Analysis* 72 (2021), page 102139.

[27] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. “Generative adversarial nets”. In: *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*. NIPS’14. Montreal, Canada: MIT Press, 2014, 2672–2680.

[28] Y. Fu, Y. Lei, T. Wang, K. Higgins, J. D. Bradley, W. J. Curran, T. Liu, and X. Yang. “LungRegNet: An unsupervised deformable image registration method for 4D-CT lung”. In: *Medical Physics* 47.4 (2020), pages 1763–1774.

[29] L. Hansen and M. P. Heinrich. “GraphRegNet: Deep graph regularisation networks on sparse keypoints for dense registration of 3D lung CTs”. In: *IEEE Transactions on Medical Imaging* 40.9 (2021), pages 2246–2257.

[30] S. Thörnqvist, J. B. Petersen, M. Høyer, L. N. Bentzen, and L. P. Muren. “Propagation of target and organ at risk contours in radiotherapy of prostate cancer using deformable image registration”. In: *Acta Oncologica* 49.7 (2010), pages 1023–1032.

[31] N. Smit, K. Lawonn, A. Kraima, M. DeRuiter, H. Sokooti, S. Bruckner, E. Eisemann, and A. Vilanova. “PelVis: Atlas-based surgical planning for oncological pelvic surgery”. In: *IEEE Transactions on Visualization and Computer Graphics* 23.1 (2017), pages 741–750.

[32] S. E. Muenzing, B. van Ginneken, M. A. Viergever, and J. P. Pluim. “DIRBoost–An algorithm for boosting deformable image registration: Application to lung CT intra-subject registration”. In: *Medical Image Analysis* 18.3 (2014), pages 449–459.

[33] H. Sokooti, G. Saygili, B. Glocker, B. P. F. Lelieveldt, and M. Staring. “Quantitative error prediction of medical image registration using regression forests”. In: *Medical Image Analysis* 56 (2019), pages 110–121.

[34] B. Glocker, N. Paragios, N. Komodakis, G. Tziritas, and N. Navab. “Optical flow estimation with uncertainties through dynamic MRFs”. In: *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2008, pages 1–8.

[35] M. Hub, M. L. Kessler, and C. P. Karger. “A stochastic approach to estimate the uncertainty involved in B-spline image registration”. In: *IEEE Transactions on Medical Imaging* 28.11 (2009), pages 1708–1716.

[36] J. Luo, S. Frisken, D. Wang, A. Golby, M. Sugiyama, and W. Wells III. “Are Registration Uncertainty and Error Monotonically Associated?” In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*. Edited by A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, and L. Joskowicz. Cham: Springer International Publishing, 2020, pages 264–274.

[37] H. Park, P. H. Bland, K. K. Brock, and C. R. Meyer. “Adaptive registration using local information measures”. In: *Medical Image Analysis* 8.4 (2004), pages 465–473.

[38] G. K. Rohde, A. Aldroubi, and B. M. Dawant. “The adaptive bases algorithm for intensity-based nonrigid image registration”. In: *IEEE Transactions on Medical Imaging* 22.11 (2003), pages 1470–1479.

[39] S. E. Muenzing, B. van Ginneken, K. Murphy, and J. P. Pluim. “Supervised quality assessment of medical image registration: Application to intra-patient CT lung registration”. In: *Medical Image Analysis* 16.8 (2012), pages 1521–1531.

[40] K. A. Eppenhof and J. P. Pluim. “Error estimation of deformable image registration of pulmonary CT scans using convolutional neural networks”. In: *Journal of Medical Imaging* 5.2 (2018), page 024003.

[41] S. Hu, L. Wei, Y. Gao, Y. Guo, G. Wu, and D. Shen. “Learning-based deformable image registration for infant MR images in the first year of life”. In: *Medical Physics* 44.1 (2017), pages 158–170.

[42] S. Miao, Z. J. Wang, and R. Liao. “A CNN regression approach for real-time 2D/3D registration”. In: *IEEE Transactions on Medical Imaging* 35.5 (2016), pages 1352–1363.

[43] X. Yang, R. Kwitt, and M. Niethammer. “Fast Predictive Image Registration”. In: *International Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*. 2016, pages 48–57.

[44] K. A. Eppenhof and J. P. Pluim. “Supervised local error estimation for nonlinear image registration using convolutional neural networks”. In: *SPIE Medical Imaging*. International Society for Optics and Photonics. 2017, 101331U–101331U.

[45] A. Dosovitskiy, P. Fischer, E. Ilg, P. Häusser, C. Hazirbas, V. Golkov, P. v. d. Smagt, D. Cremers, and T. Brox. “FlowNet: Learning optical flow with convolutional networks”. In: *IEEE International Conference on Computer Vision (ICCV)*. 2015, pages 2758–2766.

[46] K. Kamnitsas, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker. “Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation”. In: *Medical Image Analysis* 36 (2017), pages 61–78.

[47] J. Stolk, H. Putter, E. M. Bakker, S. B. Shaker, D. G. Parr, E. Piitulainen, E. W. Russi, E. Grebski, A. Dirksen, R. A. Stockley, J. H. C. Reiber, and B. C. Stoel. “Progression parameters for emphysema: a clinical investigation”. In: *Respiratory Medicine* 101.9 (2007), pages 1924–1930.

[48] K. Murphy, B. van Ginneken, S. Klein, M. Staring, B. J. de Hoop, M. A. Viergever, and J. P. Pluim. “Semi-automatic construction of reference standards for evaluation of image registration”. In: *Medical Image Analysis* 15.1 (2011), pages 71–84.

[49] Theano Development Team. "Theano: A Python framework for fast computation of mathematical expressions". In: *arXiv e-prints* abs/1605.02688 (May 2016). URL: `http://arxiv.org/abs/1605.02688`.

[50] S. Dieleman, J. Schluter, C. Raffel, E. Olson, S. K. Sonderby, D. Nouri, D. Maturana, M. Thoma, E. Battenberg, J. Kelly, et al. "Lasagne: First release". In: *Zenodo: Geneva, Switzerland* (2015).

[51] B. C. Lowekamp, D. T. Chen, L. Ibáñez, and D. Blezek. "The design of SimpleITK". In: *Frontiers in Neuroinformatics* 7 (2013), pages 1–14.

[52] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. Pluim. "`elastix`: A toolbox for intensity-based medical image registration". In: *IEEE Transactions on Medical Imaging* 29.1 (2010), pages 196–205.

[53] C. Guetter, C. Xu, F. Sauer, and J. Hornegger. "Learning based non-rigid multi-modal image registration using Kullback-Leibler divergence". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2005, pages 255–262.

[54] J. Jiang, S. Zheng, A. W. Toga, and Z. Tu. "Learning based coarse-to-fine image registration". In: *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE. 2008, pages 1–7.

[55] H. Sokooti, G. Saygili, B. Glocker, B. P. Lelieveldt, and M. Staring. "Accuracy estimation for medical image registration using regression forests". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Volume 9902. Lecture Notes in Computer Science. 2016, pages 107–115.

[56] X. Cao, J. Yang, J. Zhang, D. Nie, M. Kim, Q. Wang, and D. Shen. "Deformable image registration based on similarity-steered CNN regression". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2017, pages 300–308.

[57] M. Simonovsky, B. Gutiérrez-Becker, D. Mateus, N. Navab, and N. Komodakis. "A deep metric for multimodal registration". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2016, pages 10–18.

[58] E. Ferrante, O. Oktay, B. Glocker, and D. H. Milone. "On the adaptability of unsupervised CNN-based deformable image registration to unseen image domains". In: *International Workshop on Machine Learning in Medical Imaging*. Springer. 2018, pages 294–302.

[59] A. Sheikhjafari, M. Noga, K. Punithakumar, and N. Ray. "Unsupervised deformable image registration with fully connected generative neural network". In: (2018).

[60] A. V. Dalca, G. Balakrishnan, J. Guttag, and M. R. Sabuncu. "Unsupervised learning for fast probabilistic diffeomorphic registration". In: *arXiv preprint arXiv:1805.04605* (2018).

[61] Y. Hu, M. Modat, E. Gibson, N. Ghavami, E. Bonmati, C. M. Moore, M. Emberton, J. A. Noble, D. C. Barratt, and T. Vercauteren. “Label-driven weakly-supervised learning for multimodal deformarle image registration”. In: *Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on*. IEEE. 2018, pages 1070–1074.

[62] M.-M. Rohé, M. Datar, T. Heimann, M. Sermesant, and X. Pennec. “SVF-Net: learning deformable image registration using shape matching”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2017, pages 266–274.

[63] J. Fan, X. Cao, Z. Xue, P.-T. Yap, and D. Shen. “Adversarial similarity network for evaluating image alignment in deep learning based registration”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pages 739–746.

[64] H. Uzunova, M. Wilms, H. Handels, and J. Ehrhardt. “Training CNNs for image registration from few samples with model-based data augmentation”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2017, pages 223–231.

[65] Y. Hu, E. Gibson, N. Ghavami, E. Bonmati, C. M. Moore, M. Emberton, T. Vercauteren, J. A. Noble, and D. C. Barratt. “Adversarial deformation regularization for training image registration neural networks”. In: *arXiv preprint arXiv:1805.10665* (2018).

[66] K. Ma, J. Wang, V. Singh, B. Tamersoy, Y.-J. Chang, A. Wimmer, and T. Chen. “Multimodal image registration with deep context reinforcement learning”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2017, pages 240–248.

[67] J. Krebs, T. Mansi, H. Delingette, L. Zhang, F. C. Ghesu, S. Miao, A. K. Maier, N. Ayache, R. Liao, and A. Kamen. “Robust non-rigid registration through agent-based action learning”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2017, pages 344–352.

[68] J. O. Onieva, B. Marti-Fuster, M. P. de la Puente, and R. S. J. Estépar. “Diffeomorphic lung registration using deep CNNs and reinforced learning”. In: *Image Analysis for Moving Organ, Breast, and Thoracic Images*. Springer, 2018, pages 284–294.

[69] O. Ronneberger, P. Fischer, and T. Brox. “U-net: Convolutional networks for biomedical image segmentation”. In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, pages 234–241.

[70] S. Ioffe and C. Szegedy. “Batch normalization: Accelerating deep network training by reducing internal covariate shift”. In: *International Conference on Machine Learning*. 2015, pages 448–456.

[71] V. Nair and G. E. Hinton. “Rectified linear units improve restricted Boltzmann machines”. In: *International Conference on Machine Learning*. 2010, pages 807–814.

[72] X. Glorot and Y. Bengio. "Understanding the difficulty of training deep feedforward neural networks". In: *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. 2010, pages 249–256.

[73] M. Staring, M. E. Bakker, J. Stolk, D. P. Shamonin, J. H. Reiber, and B. C. Stoel. "Towards local progression estimation of pulmonary emphysema using CT". In: *Medical Physics* 41.2 (2014), page 021905.

[74] R. Castillo, E. Castillo, R. Guerra, V. E. Johnson, T. McPhail, A. K. Garg, and T. Guerrero. "A framework for evaluation of deformable image registration spatial accuracy using large landmark point sets". In: *Physics in Medicine and Biology* 54.7 (2009), page 1849.

[75] R. Castillo, E. Castillo, D. Fuentes, M. Ahmad, A. M. Wood, M. S. Ludwig, and T. Guerrero. "A reference dataset for deformable image registration spatial accuracy evaluation using the COPDgene study archive". In: *Physics in Medicine and Biology* 58.9 (2013), page 2861.

[76] M. Abadi et al. "TensorFlow: A system for large-scale machine learning." In: *12th USENIX Conference on Operating Systems Design and Implementation (OSDI)*. Volume 16. 2016, pages 265–283.

[77] T. Sentker, F. Madesta, and R. Werner. "GDL-FIRE 4D: deep learning-based fast 4D CT image registration". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pages 765–773.

[78] F. F. Berendsen, A. N. Kotte, M. A. Viergever, and J. P. Pluim. "Registration of organs with sliding interfaces and changing topologies". In: *Medical Imaging 2014: Image Processing*. Volume 9034. International Society for Optics and Photonics. 2014, 90340E.

[79] K. A. Eppenhof and J. P. Pluim. "Pulmonary CT registration through supervised learning with convolutional neural networks". In: *IEEE transactions on Medical Imaging* (2018).

[80] K. Murphy, B. Van Ginneken, J. M. Reinhardt, S. Kabus, K. Ding, X. Deng, K. Cao, K. Du, G. E. Christensen, V. Garcia, et al. "Evaluation of registration methods on thoracic CT: the EMPIRE10 challenge". In: *IEEE Transactions on Medical Imaging* 30.11 (2011), pages 1901–1920.

[81] M. J. Murphy, F. J. Salguero, J. V. Siebers, D. Staub, and C. Vaman. "A method to estimate the effect of deformable image registration uncertainties on daily dose mapping". In: *Medical Physics* 39.2 (2012), pages 573–580.

[82] D. Tilly, N. Tilly, and A. Ahnesjö. "Dose mapping sensitivity to deformable registration uncertainties in fractionated radiotherapy–applied to prostate proton treatments". In: *BMC Medical Physics* 13.1 (2013), page 2.

[83] C. Veiga, A. M. Lourenço, S. Mouinuddin, M. van Herk, M. Modat, S. Ourselin, G. Royle, and J. R. McClelland. "Toward adaptive radiotherapy for head and neck patients: Uncertainties in dose warping due to the choice of deformable registration algorithm". In: *Medical Physics* 42.2 (2015), pages 760–769.

[84] G. Gunay, M. H. Luu, A. Moelker, T. van Walsum, and S. Klein. "Semiautomated registration of pre- and intraoperative CT for image-guided percutaneous liver tumor ablation interventions". In: *Medical Physics* 44.7 (2017), pages 3718–3725.

[85] M. Schlachter, T. Fechter, M. Jurisic, T. Schimek-Jasch, O. Oehlke, S. Adebahr, W. Birkfellner, U. Nestle, and K. Bühler. "Visualization of deformable image registration quality using local image dissimilarity". In: *IEEE Transactions on Medical Imaging* 35.10 (2016), pages 2319–2328.

[86] J. A. Schnabel, D. Rueckert, M. Quist, J. M. Blackall, A. D. Castellano-Smith, T. Hartkens, G. P. Penney, W. A. Hall, H. Liu, C. L. Truwit, et al. "A generic framework for non-rigid registration based on non-uniform multi-level free-form deformations". In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2001*. Springer. 2001, pages 573–581.

[87] G. Saygili, M. Staring, and E. A. Hendriks. "Confidence estimation for medical image registration based on stereo confidences". In: *IEEE Transactions on Medical Imaging* 35.2 (2016), pages 539–549.

[88] D. Forsberg, Y. Rathi, S. Bouix, D. Wassermann, H. Knutsson, and C.-F. Westin. "Improving registration using multi-channel diffeomorphic demons combined with certainty maps". In: *Multimodal Brain Image Analysis*. Springer, 2011, pages 19–26.

[89] P. Risholm, F. Janoos, I. Norton, A. J. Golby, and W. M. Wells. "Bayesian characterization of uncertainty in intra-subject non-rigid registration". In: *Medical Image Analysis* 17.5 (2013), pages 538–555.

[90] I. J. Simpson, M. J. Cardoso, M. Modat, D. M. Cash, M. W. Woolrich, J. L. Andersson, J. A. Schnabel, S. Ourselin, et al. "Probabilistic non-linear registration with spatially adaptive regularisation". In: *Medical Image Analysis* 26.1 (2015), pages 203–216.

[91] J. Luo, K. Popuri, D. Cobzas, H. Ding, W. M. Wells, and M. Sugiyama. "Misdirected registration uncertainty". In: *arXiv preprint arXiv:1704.08121* (2017).

[92] R. D. Datteri and B. M. Dawant. "Automatic detection of the magnitude and spatial location of error in non-rigid registration". In: *Biomedical Image Registration*. Springer, 2012, pages 21–30.

[93] T. Gass, G. Szekely, and O. Goksel. "Consistency-based rectification of nonrigid registrations". In: *Journal of Medical Imaging* 2.1 (2015), pages 014005–014005.

[94] J. Kybic. "Bootstrap resampling for image registration uncertainty estimation without ground truth". In: *IEEE Transactions on Image Processing* 19.1 (2010), pages 64–73.

[95] M Hub and C. Karger. "Estimation of the uncertainty of elastic image registration with the Demons algorithm". In: *Physics in Medicine and Biology* 58.9 (2013), page 3023.

[96] T. Lotfi, L. Tang, S. Andrews, and G. Hamarneh. "Improving probabilistic image registration via reinforcement learning and uncertainty evaluation". In: *International Workshop on Machine Learning in Medical Imaging*. Springer. 2013, pages 187–194.

[97] S. Klein, J. P. Pluim, M. Staring, and M. A. Viergever. “Adaptive stochastic gradient descent optimisation for image registration”. In: *International Journal of Computer Vision* 81.3 (2009), pages 227–239.

[98] P. Viola and M. J. Jones. “Robust real-time face detection”. In: *International journal of computer vision* 57.2 (2004), pages 137–154.

[99] L. Breiman. “Random forests”. In: *Machine Learning* 45.1 (2001), pages 5–32.

[100] B. Glocker, D. Zikic, and D. R. Haynor. “Robust registration of longitudinal spine CT”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2014, pages 251–258.

[101] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pages 2825–2830.

[102] A. Liaw, M. Wiener, et al. “Classification and regression by RandomForest”. In: *R news* 2.3 (2002), pages 18–22.

[103] B. B. Avants, N. Tustison, and G. Song. “Advanced normalization tools (ANTS)”. In: *Insight J.* 2 (2009), pages 1–35.

[104] N. Tustison, G. Song, J. Gee, and B. Avants. *Two Greedy SyN variants for pulmonary image registration*. 2013.

[105] M. Staring, S. Klein, J. H. Reiber, W. J. Niessen, and B. C. Stoel. “Pulmonary image registration with `elastix` using a standard intensity-based algorithm”. In: *Medical Image Analysis for the Clinic: A Grand Challenge* (2010), pages 73–79.

[106] M. P. Heinrich, I. J. Simpson, B. W. Papież, M. Brady, and J. A. Schnabel. “Deformable image registration by combining uncertainty estimates from supervoxel belief propagation”. In: *Medical Image Analysis* 27 (2016), pages 57–71.

[107] S. E. Muenzing, M. Strauch, J. W. Truman, K. Bühler, A. S. Thum, and D. Merhof. “Larvalign: Aligning gene expression patterns from the larval brain of drosophila melanogaster”. In: *Neuroinformatics* 16.1 (2018), pages 65–80.

[108] I. J. Chetty and M. Rosu-Bubulac. “Deformable registration for dose accumulation”. In: *Seminars in Radiation Oncology* 29.3 (2019), pages 198–208.

[109] S. H. Cha and S. N. Srihari. “On measuring the distance between histograms”. In: *Pattern Recognition* 35.6 (2002), pages 1355–1370.

[110] T. Rohlfing. “Image similarity and tissue overlaps as surrogates for image registration accuracy: widely used but unreliable”. In: *IEEE Transactions on Medical Imaging* 31.2 (2011), pages 153–163.

[111] G. Saygili. “Predicting medical image registration error with block-matching using three orthogonal planes approach”. In: *Signal, Image and Video Processing* 14.6 (2020), pages 1099–1106.

[112] G. Gunay, S. Van Der Voort, M. H. Luu, A. Moelker, and S. Klein. “Local image registration uncertainty estimation using polynomial chaos expansions”. In: *International Workshop on Biomedical Image Registration*. Springer. 2018, pages 115–125.

[113] J. Luo, A. Sedghi, K. Popuri, D. Cobzas, M. Zhang, F. Preiswerk, M. Toews, A. Golby, M. Sugiyama, W. M. Wells, and S. Frisken. “On the applicability of registration uncertainty”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Volume 11765. Lecture Notes in Computer Science. 2019, pages 410–419.

[114] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca. “VoxelMorph: A learning framework for deformable medical image registration”. In: *IEEE Transactions on Medical Imaging* 38.8 (2019), pages 1788–1800.

[115] B. D. de Senneville, J. V. Manjón, and P. Coupé. “RegQCNET: Deep quality control for image-to-template brain MRI affine registration”. In: *Physics in Medicine and Biology* 65.22 (2020), page 225022.

[116] R. Salakhutdinov, A. Torralba, and J. Tenenbaum. “Learning to share visual appearance for multiclass object detection”. In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2011, pages 1481–1488.

[117] M. Ristin, J. Gall, M. Guillaumin, and L. Van Gool. “From categories to subcategories: large-scale image classification with partial class label refinement”. In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pages 231–239.

[118] J. Redmon and A. Farhadi. “YOLO9000: better, faster, stronger”. In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pages 7263–7271.

[119] H. Chen, S. Miao, D. Xu, G. D. Hager, and A. P. Harrison. “Deep hierarchical multi-label classification of chest X-ray images”. In: *International Conference on Medical Imaging with Deep Learning*. 2019, pages 109–120.

[120] F. Taherkhani, H. Kazemi, A. Dabouei, J. Dawson, and N. M. Nasrabadi. “A weakly supervised fine label classifier enhanced by coarse supervision”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2019, pages 6459–6468.

[121] Y. Guo, Y. Liu, E. M. Bakker, Y. Guo, and M. S. Lew. “CNN-RNN: a large-scale hierarchical image classification framework”. In: *Multimedia Tools and Applications* 77.8 (2018), pages 10251–10271.

[122] X. Shi, Z. Chen, H. Wang, D. Y. Yeung, W. k. Wong, and W. c. Woo. “Convolutional LSTM network: A machine learning approach for precipitation nowcasting”. In: *Advances in Neural Information Processing Systems*. Volume 28. 2015, pages 802–810.

[123] C. P. Tang, K. L. Chui, Y. K. Yu, Z. Zeng, and K. H. Wong. “Music genre classification using a hierarchical long short term memory (LSTM) model”. In: *International Workshop on Pattern Recognition*. Volume 10828. International Society for Optics and Photonics. SPIE, 2018, pages 334 –340.

[124] K. Simonyan and A. Zisserman. “Very deep convolutional networks for large-scale image recognition”. In: *International Conference on Learning Representations*. 2015.

[125] K. He, X. Zhang, S. Ren, and J. Sun. “Deep residual learning for image recognition”. In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pages 770–778.

[126] J. Johnson, A. Alahi, and L. Fei-Fei. “Perceptual losses for real-time style transfer and super-resolution”. In: *European Conference on Computer Vision*. Springer. 2016, pages 694–711.

[127] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio. “Transfusion: Understanding transfer learning for medical imaging”. In: *Advances in Neural Information Processing Systems*. Volume 32. 2019, pages 3342–3352.

[128] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang. “Convolutional neural networks for medical image analysis: Full training or fine tuning?” In: *IEEE Transactions on Medical Imaging* 35.5 (2016), pages 1299–1312.

[129] S. Hochreiter and J. Schmidhuber. “Long short-term memory”. In: *Neural Computation* 9.8 (1997), pages 1735–1780.

[130] D. P. Kingma and J. Ba. “Adam: A method for stochastic optimization”. In: *International conference on learning representations*. 2015.

[131] P. Mettes, E. van der Pol, and C. Snoek. “Hyperspherical Prototype Networks”. In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*. 2019, pages 1485–1495.

[132] J. M. Wolterink, A. M. Dinkla, M. H. F. Savenije, P. R. Seevinck, C. A. T. van den Berg, and I. Išgum. “Deep MR to CT synthesis using unpaired data”. In: *Simulation and Synthesis in Medical Imaging*. Edited by S. A. Tsaftaris, A. Gooya, A. F. Frangi, and J. L. Prince. Cham: Springer International Publishing, 2017, pages 14–23.

Publications

Journal articles

N. Smit, K. Lawonn, A. Kraima, M. DeRuiter, H. **Sokooti**, S. Bruckner, E. Eisemann, and A. Vilanova. “PelVis: Atlas-based surgical planning for oncological pelvic surgery”. In: *IEEE Transactions on Visualization and Computer Graphics* 23.1 (2017), pages 741–750

H. **Sokooti**, G. Saygili, B. Glocker, B. P. Lelieveldt, and M. Staring. “Quantitative error prediction of medical image registration using regression forests”. In: *Medical Image Analysis* 56 (2019), pages 110–121

M. S. Elmahdy, T. Jagt, R. T. Zinkstok, Y. Qiao, R. Shahzad, H. **Sokooti**, S. Yousefi, L. Incrocci, C. Marijnen, M. Hoogeman, and M. Staring. “Robust contour propagation using deep learning and image registration for online adaptive proton therapy of prostate cancer”. In: *Medical Physics* 46.8 (2019), pages 3329–3343

B. de Vos, F. Berendsen, M. A. Viergever, H. **Sokooti**, M. Staring, and I. Išgum. “A deep learning framework for unsupervised affine and deformable image registration”. In: *Medical Image Analysis* 52 (2019), pages 128–143

H. **Sokooti**, S. Yousefi, M. S. Elmahdy, B. P. F. Lelieveldt, and M. Staring. “Hierarchical prediction of registration misalignment using a convolutional LSTM: Application to chest CT scans”. In: *IEEE Access* 9 (2021), pages 62008–62020

S. Yousefi, H. **Sokooti**, M. S. Elmahdy, I. M. Lips, M. T. M. Shalmani, R. T. Zinkstok, F. J. W. M. Dankers, and M. Staring. *Esophageal tumor segmentation in CT images using dilated dense attention Unet (DDAUnet)*

M. S. Elmahdy, L. Beljaards, S. Yousefi, H. **Sokooti**, F. Verbeek, U. A. van der Heide, and M. Staring. *Joint registration and segmentation via multi-task learning for adaptive radiotherapy of prostate cancer*. 2021

e-print archive

H. Arjmandi-Tash, H. **Sokooti**, J. Lin, A. Kloosterman, L. M. C. Lima, and G. F. Schneider. "Biaxial compression of centimeter scale graphene on strictly 2D substrate". In: *arXiv e-prints*, arXiv:1707.07941 (July 2017), arXiv:1707.07941. arXiv: `1707.07941`

H. **Sokooti**, B. de Vos, F. Berendsen, M. Ghafoorian, S. Yousefi, B. P. Lelieveldt, I. Isgum, and M. Staring. "3D convolutional neural networks image registration based on efficient supervised learning from artificial deformations". In: *arXiv preprint arXiv:1908.10235* (2019)

S. Yousefi, H. **Sokooti**, W. M. Teeuwisse, D. F. R. Heijtel, A. J. Nederveen, M. Staring, and M. J. P. van Osch. *ASL to PET translation by a semi-supervised residual-based attention-guided convolutional neural network*. 2021. arXiv: `2103.05116 [eess.IV]`

Book chapters

B. de Vos, H. **Sokooti**, M. Staring and I. Išgum, A Frangi (Ed.). "Medical Image Analysis Text Book", Chapter "Machine Learning in Image Registration". In progress: *Medical Image Analysis*. 2021

International conference proceedings

H. **Sokooti**, G. Saygili, B. Glocker, B. P. Lelieveldt, and M. Staring. "Accuracy estimation for medical image registration using regression forests". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Volume 9902. Lecture Notes in Computer Science. 2016, pages 107–115

H. **Sokooti**, B. De Vos, F. Berendsen, B. P. Lelieveldt, I. Išgum, and M. Staring. "Nonrigid image registration using multi-scale 3D convolutional neural networks". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Volume 10433. Lecture Notes in Computer Science. 2017, pages 232–239

S. Yousefi, H. **Sokooti**, M. S. Elmahdy, F. P. Peters, M. T. M. Shalmani, R. T. Zinkstok, and M. Staring. "Esophageal gross tumor volume segmentation using a 3D convolutional neural network". In: *Medical Image Computing and Computer Assisted Intervention*. Cham: Springer International Publishing, 2018, pages 343–351

M. S. Elmahdy, T. Jagt, S. Yousefi, H. **Sokooti**, R. Zinkstok, M. Hoogeman, and M. Staring. "Evaluation of multi-metric registration for online adaptive proton therapy of prostate cancer". In: *Biomedical Image Registration*. Edited by S. Klein, M. Staring, S. Durrleman, and S. Sommer. Cham: Springer International Publishing, 2018, pages 94–104

S. Yousefi, L. Hirschler, M. van der Plas, M. S. Elmahdy, H. **Sokooti**, M. Van Osch, and M. Staring. "Fast dynamic perfusion and angiography reconstruction using an end-to-end 3D convolutional neural network". In: *International Workshop on Machine Learning for Medical Image Reconstruction*. Springer. 2019, pages 25–35

M. S. Elmahdy, J. M. Wolterink, H. **Sokooti**, I. Išgum, and M. Staring. "Adversarial optimization for joint registration and segmentation in prostate CT radiotherapy". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. Springer International Publishing, 2019, pages 366–374

Code repositories

Hessam Sokooti, RegNet, *GitHub*, version 0.2, github.com/hsokooti/RegNet

Hessam Sokooti, RegUn, *GitHub*, version 0.1, github.com/hsokooti/RegUn

Acknowledgements

The journey of this thesis was started on the 1st of April 2015 in Leiden. On that stormy day, when my umbrella broke up immediately after putting it up, it was conspicuous that this journey would not be unchallenging. Fortunately, my surrounding was full of brilliant, friendly, supportive, and scientific-enthusiastic people. They are gratefully acknowledged. I want to express my sincere gratitude to my thesis committee, opponent committee, and the (unknown) reviewers of my publications.

Firstly, I would like to thank my Ph.D. supervisors, Prof. Marius Staring and Prof. Boudewijn Lelieveldt, who offered this Ph.D. position to me. I want to give my special thanks to my mentor, friend, and daily supervisor, Marius. You were always extremely helpful, from the early debugging in C++ programming to genius, constructive and miraculous feedback on the experiments and the structure of the papers. Thank you for your overtime working before almost all submission deadlines. The freedom and support you provided for me during my Ph.D., gave me the opportunity to explore new directions with a tranquil mind.

I want to thank my cooperators, Ben Glocker for providing a broad insight into the random forests method, Gorkem Saygili for helping in the subject of registration misalignment, especially at the beginning of my Ph.D. Noeska Smit, thank you for the excellent collaboration in a practical misregistration research. Bob de Vos and Ivana Išgum, thank you for joint research in the new field of registration with deep learning. My colleagues and collaborators, Sahar Yousefi and Mohamed Elmahdy, Denis Shamonin, and Berend Stoel, thank you for all the time we spent discussing numerous challenges together. I want to thank my project co-workers from whom I received brilliant feedback: Stefan Klein, Floris Berendsen, Gokhan Gunay, Kasper Marstal, and Niels Dekker.

I would like to express my gratitude to my LKEB colleagues for making such a friendly and peaceful atmosphere. I learned a lot from you, and I received many useful feedback in Monday Morning Talks. From the old image registration group with Zhou, Yuchuan, Floris to the new deep learning group with Zhiwei, Sahar, Denis, Mohamed, Qing, Xiaowu, Jingnan, Kilany, Prerak, and Irene, all were fantastic opportunities for me to broaden my knowledge. Dear Jouke, I really enjoyed working with you in my post-doctorate position together with Labrinus. Thank Pieter for helping me with my first python program inside the MeVisLab (PyCalculator). Thank Denis for helping me with various MeVisLab components. Dear Rob, Leo, Els, Alexander, Patrick, Jeroen, Oleh, Qian, Walid, Elbaz, and Thomas, thank you for the mature-subject conversations during lunch or coffee time. Thank you Shan, I would like to thank our helpful patient secretaries, Anna-Carien, Elmi, and Helena. I enjoyed doing various sports like skiing with Antonios, Zhiwei, Kilany, and Mohamed, or swimming with Niels, Nancy, Zhiwei, and Qing. Dear Denis and Zhiwei, thank you for accepting to be my paranymphs. Your support and all of your efforts in helping me organize the defense ceremony are greatly appreciated.

The last parts of the thesis were written when I worked at Medis Medical Imaging as a researcher. Therefore, I would like to give special thanks to my colleagues, especially at the applied research group, Pieter, Viet, Hua, Evan, Nil, Jasper, Marco, Merih, Catalina, Yves, Eelco, and Mel, who makes the environment so friendly and vibrant for me in the company. I would like to express my appreciation to Prof. Hans Reiber for offering me this position.

I want to thank my family for their affectionate support.

از خانواده عزیزم که در طول مدت این دوره همیشه کنار من بودند شدیدا قدردانی می کنم، از همسر عزیزم
مونا، پدر و مادر عزیزم و خواهرم سارا به پاس محبت های بی دریغشان.

Curriculum Vitae

Hessam Sokooti was born in Iran. He received his BSc in electrical engineering from the University of Tehran in 2011. In his BSc project, he designed and implemented a two-channel electrooculography (EOG) device. He obtained his MSc degree in biomedical engineering from the K. N. Toosi University of Technology in 2014 with a master thesis about a computer-aided design (CAD) with retinal fluorescein angiography images in the diagnosis of diabetic retinopathy.

From April 2015, he started his PhD study in the Division of Image Processing (LKEB) under the Department of Radiology at Leiden University Medical Center in the Netherlands. His PhD project mainly focuses on machine learning for medical image registration.

From May 2019 to October 2019, he worked as a post-doctoral researcher in LKEB, on the project of classification of malignant and benign tissue in resected pancreatic cancer specimens. From November 2019, he started working as a researcher at Medis Medical Imaging in the Applied Research group.