UNIVERSITÄT ZU LÜBECK
INSTITUT FÜR MEDIZINISCHE INFORMATIK

From the Institute of Medical Informatics
of the University of Lübeck
Director: Prof. Dr. rer. nat. habil. Heinz Handels

Fast Image Registration for Image-guided Interventions

Schnelle Bildregistrierung für Bildgeführte Interventionen

Dissertation
for Fulfillment of
Requirements
for the Doctoral Degree
of the University of Lübeck

from the Department of Computer Sciences and Technical Engineering

Submitted by
In Young Ha
from Seoul

Lübeck, 2021

1. First referee: Prof. Dr. Mattias Heinrich
2. Second referee: Prof. Dr. rer. nat. habil. Floris Ernst

Date of oral examination: 23.08.2021

Approved for printing: 25.08.2021, Lübeck

Abstract

With the evolution of computer-aided medical procedures, especially advances in the imaging modalities, image-guidance is now widely used for interventions and clinical procedures in various anatomical sites. Image-guided interventions enable monitoring and adaptation of anatomical and/or physiological changes during clinical procedures based on realtime guidance images with minimal or even no opening up of the patients. Particularly for clinical procedures that aim to remove cancerous tissues by applying high energy to them or directly cutting them out, the aid using image-guidance can have a significant influence on the precision of the procedure and improve prognosis.

To take advantage of image-guidance, it is essential to have a fast and accurate image registration algorithm, which can align the images obtained before and during the procedure, ideally based only on the images. Planning images, which are taken for simulation of the actual procedure or for localization of tumor sites, can provide prior information on the patient's anatomy or anatomical changes. This thesis deals with data-driven fast medical image registration in the context of image-guided interventions to address the problems of computation time and lack of ground truth data. We explore different possibilities to improve computation time and registration accuracy. We start from a conventional image registration framework that can estimate dense deformation fields from sparse keypoints matching results using a statistical model from planning data. With an efficient block-matching algorithm and GPU implementation, we show that the proposed method is able to perform realtime deformable image registration with accuracy comparable to conventional registration methods. The other two methods utilize deep learning techniques that enable fast inference and training of a model without ground truth deformation data. We propose a semantic guidance for training of an end-to-end network and an instance optimization that can combine local displacement probability with global transformation conformity for improvement of registration accuracy.

Extensive evaluation of three methods has demonstrated improvement of accuracy or computation time. Moreover, all frameworks presented in this thesis can potentially deal with the registration of multi-modal image pairs and can be applied to image registration tasks in different clinical applications.

Kurzfassung

Durch die Weiterentwicklung von computerunterstützten medizinischen Anwendungen, insbesondere in der Bildgebung, ist die Bildführung bei Eingriffen und klinische Prozeduren für verschiedenen Krankheiten und anatomische Regionen heute weit verbreitet. Bildgeführte Interventionen erlauben Überwachung von und Anpassung an anatomische und physiologische Änderungen während der Durchführung von minimalinvasive Eingriffe mithilfe von echtzeitigen Bildaufnahmen. Diese Unterstützung der Behandlung kann einen signifikanten Einfluss auf die Präzision der Behandlung haben und somit die Prognose verbessern.

Um Bildführung zu nutzen, ist ein schneller und genauer Bildregistrierungsalgorithmus essentiell, der die während und vor der Behandlung erzeugten Bilder aufeinander anpassen kann. Planungsbilder, welche für die Simulation der eigentlichen Prozedur aufgenommen werden oder dazu Tumore zu orten, können im Vorfeld Informationen über die Anatomie des Patienten beziehungsweise Änderungen der Anatomie bereitstellen. Diese Arbeit behandelt datengestützte, schnelle Bildregistrierung für die Anwendung bei bildgeführten Interventionen, um die Probleme langer Rechenzeit und dem Mangel an Ground-Truth-Daten anzugehen. Es werden verschiedene Möglichkeiten zur Verbesserung der Rechenzeit und Registrierungsgenauigkeit erforscht. Die Untersuchungen beginnen mit einem konventionellen Bildregistrierungs-Framework, welches erweitert wurde um dichte Verschiebungsfelder aus den Ergebnissen von dem Matching spärlicher Keypunkte unter Nutzung eines statistischen Modells aus den Planungsdaten abschätzen kann. Mit einem effizienten Block-Matching-Algorithmus sowie einer GPU-Implementierung können wir zeigen, dass die vorgeschlagene Methode in der Lage ist, in Echtzeit eine nichtlineare Bildregistrierung durchzuführen, die in ihrer Genauigkeit vergleichbar mit konventionellen, deutlich langsameren Registrierungsmethoden ist. Die anderen beiden in dieser Arbeit entwickelte Verfahren nutzen Deep-Learning-Methoden, die ohne Ground-Truth-Deformationsdaten schnelle Inferenz und schnelles Training eines Modells ermöglichen. Wir schlagen vor, eine semantische Unterstützung für das End-to-End-Training eines Netzwerks sowie eine instanzbasierte Optimierung zu nutzen, welche lokale Verschiebungswahrscheinlichkeiten mit globaler Regularisierung der Transformation kombiniert, um die Registrierungsgenauigkeit zu verbessern.

Die ausführliche Auswertung der drei Methoden hat gezeigt, dass eine deutliche Verbesserung der Rechenzeit erreicht werden konnte. Außerdem können alle in dieser Arbeit

entwickelte Algorithmen potenziell genutzt werden, um multi-modale Bildpaare zu registrieren und können für Registrierungszwecke in verschiedenen klinischen Anwendungen genutzt werden.

Contents

Chapter 1

Introduction

1.1 Motivation

In recent years, radiological images such as ultrasound (US), computed tomography (CT), or magnetic resonance imaging (MRI) are increasingly acquired and used in various phases of clinical procedures, not only limited to diagnostic purposes but also for planning, guidance, and evaluation of treatment and monitoring of disease progression. Most of the time, images are acquired from patients at different time points and often with the use of different imaging modalities. Therefore, it is essential to be able to compare these images to obtain the desired information. With recent advances in computer-assisted interventions, it is possible to compare and/or fuse images quantitatively and directly using computational image registration algorithms. Interests in development of accurate, robust, and particularly fast image registration algorithms are rising with the increasing application of image-guided interventions in clinical practice for minimal or non-invasive procedures.

In the past, most research on image registration approaches was focused on the improvement of accuracy and robustness, whereas the reduction of computation time often was not the primary concern. In the last few decades, computational capacity has improved rapidly, and different image processing techniques were developed for efficient computing. Advances in graphics processing unit (GPU) programming and machine/deep learning approaches have opened up the possibility of realtime image processing. Furthermore, the development and advancement of various image-guided interventions encourage the research and application of such algorithms, as they provide abundant radiological images that are continuously acquired during treatment for guidance, which would make time-consuming algorithms infeasible.

Image-guided interventions are applied in many different clinical procedures due to their advantage to measure and visualize the internal structures of patients without opening them up during procedure. Clinical procedures such as neurosurgery, where it is difficult to identify structures even after opening the skull, or radiation therapy,

which is usually done without opening the patient, are the ones for which in particular image-guidance can provide enormous advantages. For these applications, it is essential to monitor changes in the shape or location of the region of interest between different patient care phases (e.g. planning, treatment, and assessment) or during treatment. This intends to minimize possible damage caused by such changes.

Motivated by this, the main purpose of this thesis was to develop and evaluate fast model-based or learning-based image registration frameworks for medical images. The focus was particularly on efficient computation and the use of a (statistical or learned) model, that can be generated based on existing data or via self-supervision. Various possibilities for acceleration of computation time are explored including GPU programming, modification of existing algorithms for efficient computing, and reduction of model complexity via network pruning.

The presented works can be divided into two parts. In the first part, we focus on the speed-up of a conventional image registration framework. In the second part, we investigate different possibilities to improve accuracy and robustness of image registration frameworks based on deep learning, which inherently guarantees fast computation. Therefore, we explore possible advantages of incorporating auxiliary data into an end-to-end learning approach. Following the weakly-supervised approach, an unsupervised approach is also presented, which can be seen as preliminary work for an iterative optimization scheme that can be further extended to be integrated into an end-to-end learning framework.

1.2 Overview and Contribution

In the following, an overview of this thesis will be given, with the main contributions highlighted in the text.

> In **Chapter 2** "Background", important background information about image registration in medical image analysis will be provided for a better understanding of this thesis. Basic terms used in the thesis and general ideas about image registration will be explained, and some relevant image registration approaches will be described in detail. In addition, evaluation metrics for quantitative and qualitative assessment of the proposed and compared image registration approaches will be shortly explained. As for the medical background, a brief summary and examples for clinical applications of image-guided interventions will be given.

Chapter 3 to Chapter 5 comprise the method part of this thesis, which can be divided into two parts. In the first part, a statistical model-based approach is presented. This

approach requires an explicit knowledge for model generation and focuses on speeding up the computation time, while preserving the accuracy of a state-of-the-art registration approach.

Chapter 3 "Model-based Sparse-to-Dense Deformable Image Registration" presents **a statistical model-based realtime registration for online motion adaptation scenario in image-guided radiation therapy**, particularly with MRI and US guidance. A discrete deformable image registration approach was developed that incorporates patient-specific information from the pre-treatment phase into an online respiratory motion adaptation. An existing state-of-the-art registration algorithm was required to generate a statistical respiratory motion model of a patient. Then the generated statistical model is used in combination with a block-matching and a novel coupled convex optimization algorithm to estimate a deformation field. For acceleration of computation time, **a discrete registration algorithm was implemented for a sparse set of keypoints on GPU and a novel approach for efficient block-matching algorithm was developed**.

The second part of this thesis involves the development and evaluation of registration approaches using deep learning frameworks. In Chapter 4 and 5, deep learning-based approaches are presented, where the model is generated only based on the given image data or additionally with auxiliary data. With deep learning approaches, fast computation in inference time can be readily achieved, whereas the accuracy and robustness of the approaches should be evaluated carefully.

Chapter 4 "Weakly-Supervised Image Registration using Segmentation for Large Deformation" presents **a deep learning-based end-to-end approach that trains the model with weak supervision via auxiliary label data**. With extensive experiments, we evaluate the influence of each component in the proposed framework. Moreover, we show that accuracy and robustness of the learned model can be improved by indirect supervision using semantic information given by expert segmentations.

Chapter 5 "Discrete multi-modal registration for image-guided surgery using deep feature learning and instance optimization" presents **a discrete image registration framework that combines a learned model for feature extraction with a conventional registration approach** based on a discrete block-matching algorithm. The efficient block-matching algorithm developed in the first part of the work (Chapter 3) is used again for fast

computation. For regularization of estimated deformation, we also present **a novel instance optimization algorithm**, which enables optimization of estimated displacement vectors based on local dissimilarity probabilities and global transformation conformity. Moreover, network pruning was performed to reduce the complexity of the learned model for feature extraction.

Finally, **Chapter 6** "Conclusion and outlook" summarizes important aspects presented in this thesis and suggests further research prospects regarding remaining problems and challenges.

All novel approaches presented in this thesis are published in well-respected international journals after going through an extensive peer-review process.

1.3 List of Own Publication

This thesis is based on the following scientific articles, which has been published as first author after going through a peer-review process:

- Ha, I. Y., Wilms, M., and Heinrich, M. (2020). Semantically guided large deformation estimation with deep networks. *Sensors*, 20(5):1392
- Ha, I. Y., Hansen, L., Wilms, M., and Heinrich, M. P. (2019). Geometric deep learning and heatmap prediction for large deformation registration of abdominal and thoracic CT
- Ha, I. Y. and Heinrich, M. P. (2019). Comparing deep learning strategies and attention mechanisms of discrete registration for multimodal image-guided interventions. In *Large-Scale Annotation of Biomedical Data and Expert Label Synthesis and Hardware Aware Learning for Medical Imaging and Computer Assisted Intervention*, pages 145–151. Springer
- Ha, I. Y., Wilms, M., Handels, H., and Heinrich, M. P. (2018). Model-based sparse-to-dense image registration for realtime respiratory motion estimation in image-guided interventions. *IEEE Transactions on Biomedical Engineering*, 66(2):302–310
- Ha, I. Y. and Heinrich, M. P. (2021). Modality-agnostic self-supervised deep feature learning and fast instance optimisation for multimodal fusion in ultrasound-guided interventions. *Computer Methods and Programs in Biomedicine*, 211:106374

Other publications related to this work, but without being the first author:

- Wilms, M., Ha, I. Y., Handels, H., and Heinrich, M. P. (2016). Model-based regularisation for respiratory motion estimation with sparse features in image-guided interventions. In Ourselin, S., Joskowicz, L., Sabuncu, M. R., Unal, G., and Wells, W., editors, *19th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) 2016*, volume 9902 of *Image Processing, Computer Vision, Pattern Recognition, and Graphics*, pages 89–97, Athen. Springer International Publishing, Springer International Publishing

Chapter 2

Background

Image registration is an image processing technique that aims to align (spatially map) two images to enable comparison between structures of interest in the input images. To achieve this, an optimal spatial transformation has to be determined, which minimizes (or maximizes) the dissimilarity (or similarity) between two images.

Since the introduction of digital imaging in the 1970s, computational approaches for medical image analysis have been investigated. Until the 1990s, medical image analysis was performed via low-level pixel processing (e.g. edges or line detectors and algorithms like region growing) and mathematical modeling (e.g. fitting lines, circles and ellipses). In the end of the 1990s, approaches using supervised techniques became popular that incorporated prior knowledge from training data to develop a model. Techniques based on statistical modelling such as active shape models (ASMs) were proposed, atlas based methods and the concept of feature extraction was introduced. Starting around 2014, deep learning based algorithms were being actively investigated [Litjens et al., 2017]. Since the success of the AlexNet [Krizhevsky et al., 2017], the application of convolutional neural networks (CNNs) for image analysis has become increasingly popular in computer vision and medical image analysis and the research on medical image registration using deep learning has also increased drastically in the last few years [Fu et al., 2020].

At the same time, the number of acquired medical images throughout the clinical process for diagnosis or treatment has been increased with improvements of different medical imaging systems. With these improved systems, a faster acquisition of high quality medical images became possible, which also enabled image-guidance in many different interventions. As a result, there is an increasing need for automatically comparing or fusing these images that are taken at different times, by different imaging modalities, or in different coordinate systems to enable clinicians to obtain relevant information from these images. Thus, image registration is at the core of many applications. It enables combination of different information provided by different imaging modalities (multi-modal image fusion), the investigation of spatial or temporal changes in patient's anatomy (longitudinal studies), and the use of population modelling or statistical atlas generation.

Digital images consist of discrete pixels or voxels, each with an assigned intensity based on the structure they belong to and can be represented as a mapping function:

$$I : \mathbf{x} \in \Omega \mapsto \mathbb{R} \tag{2.1}$$

where $\mathbf{x} \in \Omega$ is a point (pixel/voxel) in an image domain Ω. For simplicity, we consider an image registration between two images (groupwise registration would enable the alignment of three or more images). Given two images, we refer to the image that stays unchanged as *fixed*, *reference* or *target* image (denote as I_F or I_R) and the image that is transformed using the estimated transformation parameters (or a deformation field) in order to be aligned with the fixed image as *moving*, *template* or *source* image (I_M or I_T). A spatial transformation $\mathcal{T}$ transforms image points of the moving image, so that the corresponding pixels are aligned with the fixed image. Usually, *backward warping* is performed, which maps each point $\mathbf{x}$ of the warped moving image to the corresponding point of the fixed image $\mathcal{T}(\mathbf{x})$. This enables simpler computation of the mapping, since the image intensity of non-voxel locations can be calculated through interpolation of the neighboring voxels. The goal of each registration problem is to find an optimal transformation that minimizes an objective function (or energy function) $\mathcal{E}$:

$$\mathcal{E} = \mathcal{D}(I_F, \mathcal{T}(I_M)) + \lambda \mathcal{R}(\mathcal{T}) \tag{2.2}$$

The objective function for image registration typically consists of a (dis)similarity term $\mathcal{D}$ that measures the (dis)similarity between the images and a regularization function $\mathcal{R}$ that seeks to deal with the ill-posedness of the image registration problem.

To find a solution for the objective function and to determine an optimal transformation to align two images, the following three components have to be considered: a deformation model, a matching criterion and an optimization method [Sotiras et al., 2013].

In this chapter, general procedures of classical image registration approaches will be explained with the focus on the methods relevant to this thesis (section 2.1) as well as deep learning techniques focusing on medical image registration to give a brief introduction to the background of our own contributions (section 2.2). After each of these sections, some important state-of-the-art image registration approaches will be introduced. In addition, evaluation metrics for quantitative and/or qualitative evaluation of image registration results will be shortly addressed (section 2.3). Finally, a general introduction on image-guided interventions with some of the examples of medical/clinical use of registration will be given in section 2.4.

2.1 Classic Image Registration

Image registration approaches, which do not rely on machine/deep learning techniques, commonly referred to as *classic (or conventional) image registration* methods. The definition, classification and characteristics of various classic medical image registration approaches are well explained in several review articles [Maintz and Viergever, 1998; Hill et al., 2001; Fischer and Modersitzki, 2008; Modersitzki, 2004; Maurer and Fitzpatrick, 1993; Sotiras et al., 2013]. Here, we summarize some key elements for this thesis.

The main components of classic image registration can be categorized into a matching criterion, a deformation model and an optimization algorithm [Sotiras et al., 2013]. The matching criterion can evaluate how well the images are aligned and is represented by the similarity cost term in the objective function. The optimal transformation should maximize similarity or minimize dissimilarity between the images after applying the transformation. The matching can be performed between a set of landmarks that are found on salient locations in the image and often corresponds to meaningful anatomical structures. The drawbacks of methods using this type of matching are the reliability of selected landmarks and the reduction of accuracy in interpolated deformation fields due to the sometimes large distances between the landmarks, when a deformable transformation model is used. Instead of finding matches between landmarks, or in addition to it, intensity information can be incorporated into the matching term. In this case, using an appropriate similarity measure is of significant importance. For mono-modal image registration, similarity measures computed directly based on image intensity, such as the sum of absolute differences (SAD), the sum of squared distances (SSD), or the cross-correlation (CC) might be used. However, for multi-modal image registration, similarity measures based on information theory, such as mutual information (MI), or structural feature descriptors are often used in practice.

Based on how the geometric constraints of the actual deformation between two images, can be formalized to find an optimal mapping or alignment, a deformation model such as rigid, affine, or deformable should be determined. A rigid transformation model can represent translation and rotation of an image using 6 parameters (in 3D), which can be adequate for registration of rigid structures such as bones or brain in a closed skull. An affine transformation model can additionally deal with scaling and shearing using 12 parameters. Both rigid and affine transformation models usually transform the image globally, and typically cannot capture local changes. However, a deformable registration model is local, non-linear and potentially non-parametric and can have up to millions of degrees-of-freedom. When a deformable alignment is estimated, the output is usually a dense deformation field, where a displacement vector is determined for each voxel or control point (in the case of B-spline transformations [Rueckert et al., 1999]).

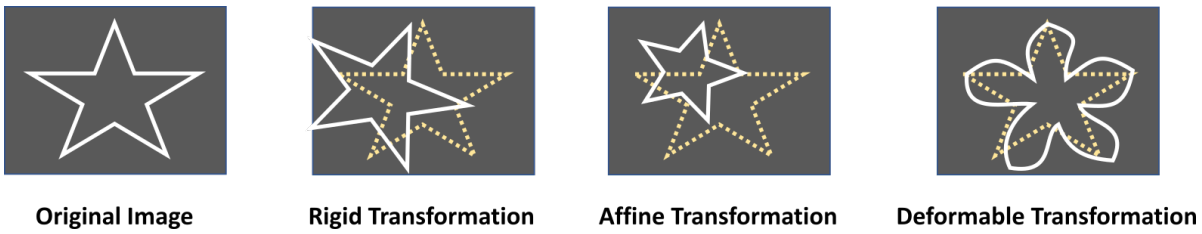


Figure 2.1: Types of transformation in 2D example images.

Deformable registration has a wide variety of applications in medical imaging, due to local deformations of the organs in the abdominal region and organs with motion (e.g. heart and lung). Moreover, tissue changes or tumor growth can only be represented with local deformation.

Figure 2.1 shows the different types of transformation. A rigid transformation constrains the deformation between two images to be parameterized only by translation and rotation. With an affine transformation, scaling and shearing can be compensated in addition to translation and rotation. Both, rigid and affine transformations, can be represented with 6 parameters in 2D, however a rigid transformation is only 3 degrees of freedom. Deformable transformations or nonlinear transformations, however, are usually represented with more parameters defined on a coarse set of control points or even on image voxels, with the latter resulting in a dense deformation field.

Optimization methods are used to infer optimal transformations by finding a (global) minimum of the objective function that consists of a similarity cost term and a regularization term. Depending on the nature of the variables, either a continuous or a discrete method can be used. Here, a brief explanation of conventional image registration methods will be given, focusing on a discrete optimization method combined with structural feature-based similarity measures for a deformable image registration. A concrete state-of-the-art algorithm based on this configuration is also explained in detail in section 2.1.3.

In the following, we will now briefly define concrete examples and mathematical models of the different components of a registration by exemplary representations, starting with matching criteria.

2.1.1 Matching criteria

2.1.1.1 Image intensity matching

The simplest similarity measure is the sum of squared differences (SSD) that computes the similarity based on intensity values of two images directly. The similarity is calcu-

lated by adding up the difference of image intensities between voxels at corresponding positions $\mathbf{x}$:

$$SSD(I_1, I_2) = \sum_{\mathbf{x}\in\Omega} (I_1(\mathbf{x}) - I_2(\mathbf{x}))^2 \tag{2.3}$$

where $I_1, I_2 \in \Omega$ are images with image domain Ω and $\mathbf{x}$ is a voxel. Although it is simple and efficient to compute similarity using SSD, it has several limitations. The images to be compared have to be from the same modality, i.e. have a similar intensity distribution. It is also sensitive to image noise and image artefacts.

Another way to compare image similarity based on image intensity is using the normalized cross correlation (NCC) between two images. It is much more robust to image noise than SSD, since it assumes and allows for a linear relationship between the intensity values in the images. NCC is computed as:

$$NCC(I_1, I_2) = \sum_{\mathbf{x}\in\Omega} \hat{I}_1(\mathbf{x})\hat{I}_2(\mathbf{x}), \quad \text{with} \quad \hat{I} = \frac{I - \bar{I}}{\sqrt{\sum (I - \bar{I})^2}} \tag{2.4}$$

where $\hat{I}$ is the intensity normalized image and $\bar{I}$ is the mean intensity of the image. NCC can either be computed globally, with Ω being the image domain, or for overlapping local patches.

2.1.1.2 Feature-based matching

When simple similarity measures are not applicable, one way to perform image registration is based on transforming the input images beforehand. To this end feature representations are used. Feature based registration approaches use feature descriptors to indirectly compute similarity between input images. A feature descriptor characterizes local appearance around each position in the input image with a series of values (represented as a feature vector) so that a descriptor can be distinguished from another. The description should (ideally) be invariant under image transformation such as translation, rotation, and scaling. Prominent feature descriptors are e.g. SIFT (scale invariant feature transform) [Lowe, 2004], SURF (speed up robust feature) [Bay et al., 2006] and BRIEF (binary robust independent elementary features) [Calonder et al., 2010]. In this thesis, we use a state-of-the-art local structure descriptor that is suitable for both mono-modal and multi-modal image registration, called *MIND/SSC descriptor*.

Modality Independent Neighborhood Descriptor (MIND) with Self Similarity Context (SSC) The MIND descriptor is a feature descriptor proposed in Heinrich et al. [2012a] for multi-modal image registration. It is based on the assumption that local image structures can be represented by the similarity of small neighboring image

patches within one modality, which can be applied to each modality independently and leads to similar descriptors for the same geometric primitives. This enables a comparison between images of different modalities using an SSD dissimilarity. The MIND descriptor is computed based on a distance D_p, a variance estimate V and spatial search region R and defined as:

$$\text{MIND}(I, \mathbf{x}, \mathbf{r}) = \frac{1}{n}\exp(-\frac{D_p(I, \mathbf{x}, \mathbf{x}+\mathbf{r})}{V(I, \mathbf{x})}), \quad \mathbf{r} \in R \tag{2.5}$$

where $\mathbf{x}$ is a location (pixel or voxel) on the image, $\mathbf{r} \in R$ is a location in the search region and n is a normalization factor that sets the maximum descriptor value to 1. The resulting MIND descriptor is then represented as a multidimensional vector, with a vector size of $|R|$.

The feature distance $D_p(I, \mathbf{x}_1, \mathbf{x}_2)$ of two image voxels $\mathbf{x}_1$ and $\mathbf{x}_2$ is computed as the SSD between all positions in the two image patches, centered at $\mathbf{x}_1$ and $\mathbf{x}_2$, with $\mathbf{x}_2$ in the search region. The search region R is limited to the six-neighborhood, which enables efficient computation. The variance estimate V is calculated as the mean of the patch distances within a six-neighborhood. The resulting MIND descriptor is, therefore, a vector with six values.

In Heinrich et al. [2013b], the MIND descriptor with self-similarity context is introduced (MIND/SSC descriptor), which is significantly less influenced by the noise in a local neighborhood, making it robust for feature description even of US images. The MIND/SSC descriptor computes similarity among neighboring patches around the voxel of interest instead of computing the similarity between the considered patch (centered at the voxel of interest) with its neighboring patches. MIND/SSC descriptors focus on finding context around the voxel of interest, while the MIND descriptor extracts representations of local shape or geometry. Example feature maps generated using the SSC descriptor are shown in Figure 2.2.

2.1.2 Determining optimal transformation

2.1.2.1 Regularization

Having image representations at hand, the next step to perform a registration is to determine how to compute an optimal transformation. Image regularization is, in general, an ill-posed problem, i.e. there is no unique solution for an optimal transformation. To deal with this problem, a regularization is used that provides an additional constraint to solve the problem. Regularization aims to penalize or favor a certain property of the transformation, and in the case of a non-parametric deformation mode, it dictates the nature of the transformation [Sotiras et al., 2013]. This usually involves applying

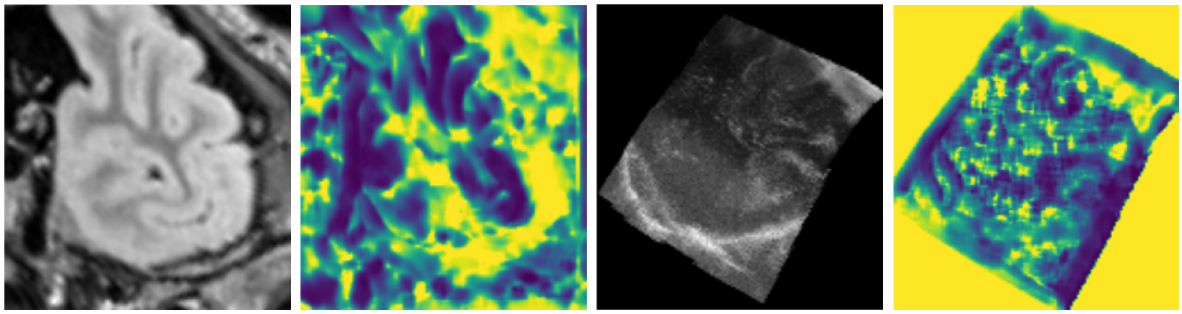


Figure 2.2: Example feature maps generated using the SSC descriptor. The first and the third images are the original images of different imaging modalities (MR and US); the feature maps generated using the SSC descriptors are shown to the right of each original image.

prior knowledge regarding the nature of the deformation present in the data (physical properties of underlying anatomical structures) or certain properties of the transformation, and helps to avoid local minima during optimization. Some important properties of transformation for medical applications are topology preservation, diffeomorphism, symmetry, and inverse consistency. Topology preservation is one of the most important properties for medical image registration and can be evaluated with the Jacobian determinant of the deformation field. Topology is also preserved when the transformation is diffeomorphic. A diffeomorphic transformation is also invertible, and both the function and its inverse are differentiable. However, it does not guarantee inverse consistency or symmetry.

Regularization of transformation can be assumed by using specific transformation models which are derived from physical models (e.g. diffusion models), from interpolation theory (e.g. free form deformations) or by using knowledge-based deformation models (e.g. statistical model). One of the regularization methods based on interpolation, which is widely used for medical image registration, is the B-spline transformation model [Rueckert et al., 1999]. The transformation space is restricted by the parametrization of transformations into lower degrees-of-freedom and the resulting transformation is usually smooth, whereas the preservation of topology is not guaranteed [Rueckert et al., 2006]. Knowledge-based regularization can be used when registration is performed for a fixed target image or for specific anatomical organs. Registration for motion estimation of a specific organ is an example of such case, where statistical model based regularization is often used [McClelland et al., 2013; King et al., 2012; Klinder and Lorenz, 2012; Boye et al., 2013; Preiswerk et al., 2014; Stemkens et al., 2016].

Statistical model based regularization When the domain of the reference image is fixed, i.e. in the case where multiple template images have to be registered to a reference image, a high dimensional statistical deformation model can be generated. One example application is the intra-interventional respiratory motion estimation in image-guided radiation therapy, where images from the planning phase can be used to generate a statistical motion model. The ideas presented in Chapter 3 are based on this. The generated motion model is assumed to contain information on the respiratory motion pattern of the patient and can be used to penalize the motion estimation that diverges from it. The data used to generate (learn) the statistical model should be representative of the deformation variations that will be present in the new data, as a prerequisite for the method to work.

One way to generate statistical models is performing a Principal Component Analysis (PCA), which enables a dimensionality reduction. Any transformation in the model space can be represented as a linear combination of few principal components with appropriate coefficients, and the inferred transformation is regularized by a restricted transformation space.

2.1.2.2 Optimization

Now that we know what dissimilarity measure, regularization, and transformations are, we can move on to the actual optimization process. Choosing an appropriate optimization method is important, since it directly influences the quality of the registration result. Optimization methods can be categorized into continuous and discrete optimization, based on the nature of the variables they deduce. The variables derived by continuous optimization methods have real values, whereas the discrete optimization infers variables take discrete values. Continuous optimization only works for differentiable objective functions. The optimization itself is performed by minimizing the objective function $\mathcal{E}$ (Equation 2.2). $\mathcal{E}$ is parameterized by a transformation parameter θ and their according update rule is as follows:

$$\theta_{t+1} = \theta_t + \alpha_t \mathbf{g}(\theta_t) \tag{2.6}$$

where θ denotes the vector of the transformation parameters, t is the iteration index, α is the step size and $\mathbf{g}$ is the search direction.

Another class of optimization techniques, which can also be used for non-differentiable functions, is called discrete optimization. Among others, Markov Random Fields (MRF) belongs to this class, which formulates the optimization problem as a probabilistic graphical model represented by an undirected graph. In the following, we give a brief summary for two popular optimization methods: stochastic gradient descent (continuous) and belief-propagation (discrete).

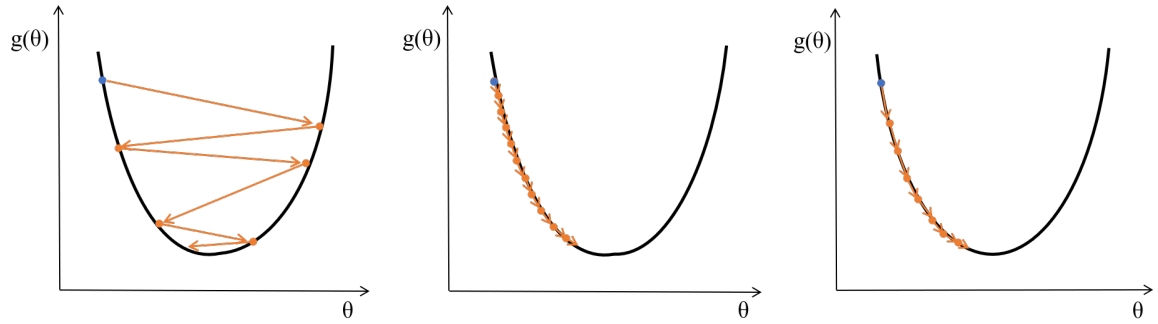


Figure 2.3: Illustration of convergence in stochastic gradient descent with different step size. Too large step size causes oscillation (left) and too small step size result in a slow convergence (middle). Step size can be decreased in each step to facilitate convergence (right).

(Stochastic) gradient descent (SGD) optimization Gradient descent optimization minimizes the objective function parameterized by the transformation parameters $\mathcal{E}(\theta)$ by following the negative gradient direction, i.e. $\mathbf{g}(\theta_t)$ of the equation for updating rule (Equation 2.6) is equivalent to $-\nabla_{\theta_t}\mathcal{E}(\theta_t)$. It is often used in the large deformation diffeomorphic metric mapping (LDDMM) frameworks and also in the free-form deformation (FFD) registration method of Rueckert et al. [1999].

One of the variations of gradient descent is stochastic gradient descent optimization, which updates the parameters based on an approximation of the gradient. It is often used when computation of the gradient is difficult. Compared to deterministic gradient methods such as gradient descent, it can be computationally efficient for each iteration, but the convergence may be slower. It is applied for global linear or cubic B-spline FFD transformations, where the deformation models have lower degrees of freedom. Moreover, it is also the most popular optimization algorithm used for training of neural networks. In a deep learning framework, the stochastic gradient descent optimization performs a parameter update for a subset of training example and thereby the objective function can strongly oscillate (Figure 2.3). However, convergence can be facilitated by decreasing the step size (equivalent to the learning rate in deep learning framework) with time. Klein et al. [2009] also introduced an adaptive stochastic gradient descent (ASGD) optimization method, which is a SGD for adaptive step size prediction.

Belief-propagation Belief-propagation is a message passing algorithm that performs inference based on local message exchange between the nodes in graphical models. For an image registration, each control point of the deformation field is considered as a node, which is connected with its neighboring nodes by edges. An optimal discrete label

for each node is sought by minimizing the energy function consisting of data cost and regularization cost. Efficient belief-propagation algorithms are proposed by Yang et al. [2010] and Felzenszwalb and Huttenlocher [2006] and applied for respiratory motion estimation by Heinrich et al. [2012b]. The dense image registration method proposed by Heinrich et al. [2012b] that uses this optimization algorithm will be introduced in the following, which is an example for a state-of-the-art classical image registration method.

2.1.3 Dense displacement sampling (Deeds)

Dense displacement sampling (Deeds) is a state-of-the-art classic deformable image registration method introduced by Heinrich et al. [2012b] and further extended in Heinrich et al. [2013a]. The main idea of the method is to use MRF-based discrete optimization to minimize the cost function of a deformable image registration. Using a discrete optimization is beneficial for medical image registration, especially when there are local motion discontinuities such as sliding motion. The method uses a minimum spanning tree (MST) to represent the underlying anatomical structures instead of using a fully connected image grid, thereby avoiding the assumption of all neighboring B-spline nodes (i.e. control points) to have similar motion. In addition, with the sparsely connected control points (nodes) on the image grid, the dimensionality problem can be addressed, which often is the limiting factor in discrete optimizations for deformable image registration. The original evaluation of the method is performed on 3D Lung CT datasets (DIR-Lab) for intra-patient registration of inhale and exhale images, which has shown a significantly improved accuracy compared to the other methods using discrete optimization [Heinrich et al., 2012b].

As introduced in section 2.1 above, the aim of a discrete registration method is to assign motion vectors to each voxel to transform the image, i.e. to find an optimal deformation field. For this, first a graph is defined on an image grid as depicted in Figure 2.4 on the left, where the voxels or groups of voxels correspond to the nodes of the graph. A set of labels are defined and for each node $p \in \mathcal{P}$. Each label f_p of a node p corresponds to a three-dimensional displacement vector $\mathbf{u}_p \in \{u_p, v_p, w_p\}$ within the search region. An optimal label configuration is to be determined that minimizes a combined cost function. The cost function usually consists of a unary term (or data term) and a regularization cost term. The data term is computed between a voxel of one image and the displaced voxels of the other image to measure the similarity between them, and can be determined independently for each node. Unlike other methods that use a multi-resolution approach, the data term of the *deeds* method is computed in the original image resolution. The regularization term is computed pair-wise between a node

and its directly connected neighboring nodes. It penalizes the deviation of displacements from the neighboring voxels, thereby ensuring a globally smooth transformation.

Different metrics can be used as similarity measure $\mathcal{D}$. In Heinrich et al. [2012b], SAD of image intensities is used as similarity measure, whereas other similarity measures, such as SAD of intensity-invariant SSC descriptors [Heinrich et al., 2013b] can also be used for the similarity cost calculation [Heinrich et al., 2014]. Using an SSC descriptor enables the registration of multi-modal images and is more robust. Therefore, the version of deeds with SSD descriptor is used as the gold standard method in Chapter 3.

For the plausibility of the estimated deformation field, the labels of neighboring nodes should be coherent. For the *deeds* method, a diffusion regularization or total variation can be used as pair-wise regularization term, for which the simplification of computational complexity using min-convolution [Felzenszwalb and Huttenlocher, 2006] is possible. The regularization term penalizes the label difference between two neighboring nodes p, q with respect to their Euclidean distance:

$$\mathcal{R}(f_p, f_q) = \mathcal{R}(\mathbf{u}_p, \mathbf{u}_q) = \frac{||\mathbf{u}_p - \mathbf{u}_q||^2}{||\mathbf{x}_p - \mathbf{x}_q||} \tag{2.7}$$

where $\mathbf{u}$ is the displacement vector and $\mathbf{x}$ is the coordinate vector (spatial location) of the node.

The energy term is then defined as:

$$E(f) = \sum_{p \in \mathcal{P}} \mathcal{D}(f_p) + \alpha \sum_{p,q \in \mathcal{N}} \mathcal{R}(f_p, f_q) \tag{2.8}$$

which is optimized by message passing (belief propagation, see also above) on a minimum spanning tree using dynamic programming.

To generate a minimum spanning tree, only the edge with minimum total edge cost is left and the remaining edges connecting a node to its neighbors is removed. The edge weight is computed as the similarity between a node and its neighboring node, i.e. in the multi-level approach, the similarity between a group of voxels. When the edge costs are determined, a minimum spanning tree can be generated using Prim's algorithm [Prim, 1957], which returns the sorted list of all nodes with the index of their parent node. The optimal label (i.e. displacement vector) for each node p is then determined by computing the cost C_p, given the displacement label f_q of the parent node q:

$$C_p(f_q) = \min_{f_q} \Big(\mathcal{D}(f_q) + \alpha \mathcal{R}(f_p, f_q) + \sum_c C_c(f_p) \Big) \tag{2.9}$$

where c are the children of the current node p. For a leaf node, which has no children, Equation 2.9 can be directly evaluated and by changing min to arg min the best displacement vector can be found for each node.

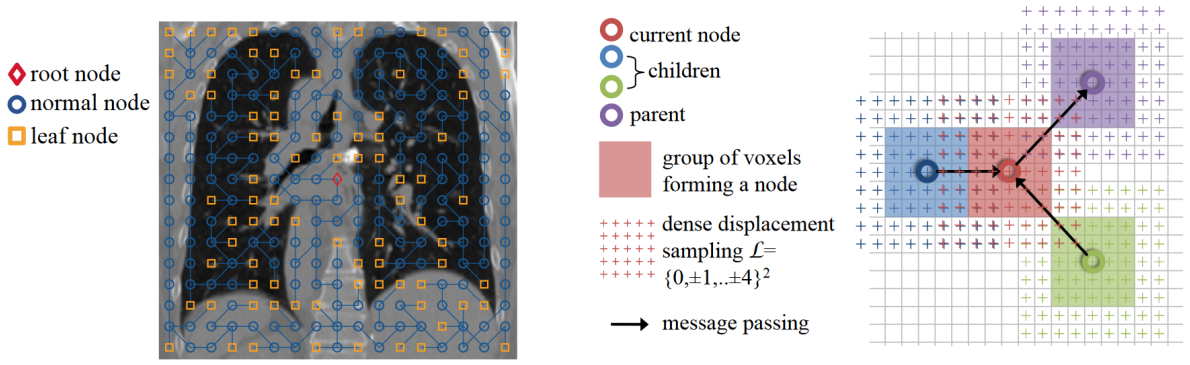


Figure 2.4: Example of a minimum spanning tree of a 2D coronal slice of a lung CT (left) and schematic description of discrete optimization scheme using dense displacement sampling (right). Images from Heinrich et al. [2012b].

Usually, for a deformable registration, one image is chosen to be the fixed image and the other as the moving image. The estimated deformation transforms the moving image to fit the fixed image. However, this may result in an inverse consistency error, which is the difference between AB and the identity, where A and B are the transformations from moving image to fixed image (forward) and the transformation from fixed to moving image (backward), respectively. As mentioned in section 2.1.2, topology preservation is one of the most important properties in medical image registration and can be achieved by minimizing the singularities in the displacement field. To minimize the singularities and bias present in the estimated transformation, when only one directional transformation is considered, an additional step is performed in the *deeds* method to ensure the symmetry. Similar to the symmetric deformable registration used in Avants et al. [2008], the forward transform A and backward transform B are computed independently. Then half-length inverse transforms $A^{-1}(0.5)$ and $B^{-1}(0.5)$ are calculated using a fast iterative inversion [Chen et al., 2008], which ensures the new symmetric transforms $A^s = A(0.5) \circ B^{-1}(0.5)$ and $B^s = B(0.5) \circ A^{-1}(0.5)$ to be inverse consistent.

The whole process is repeated in several levels (multi-level approach). For each level, the image is subdivided into non-overlapping groups of voxels and for each group, a node (or control point) is defined as the center of a group of voxels. The similarity cost (unary term) is computed for each voxel in the original image resolution, and the similarity cost for a node at each level is calculated by aggregating the similarity cost of all voxels in the same group. The regularization term is also computed for each group of voxels, since it is computed based on the label and the spatial location of each node. The displacement field from the previous level is upsampled for the next level and used as the initial deformation at the next level. With increasing levels, the number of nodes

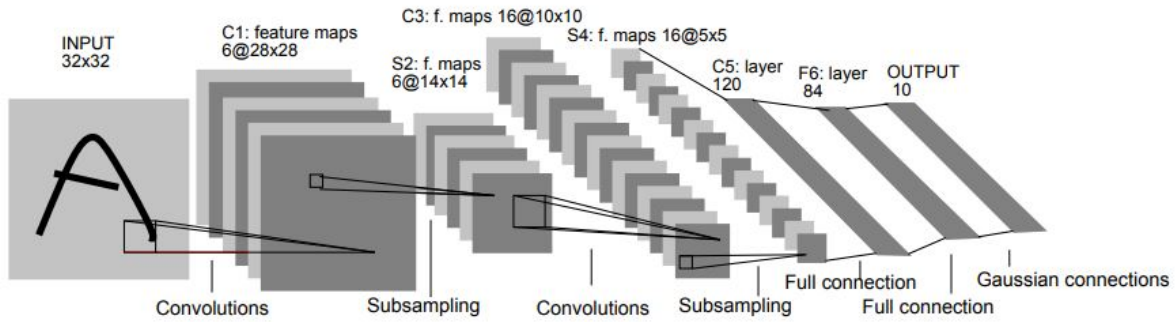


Figure 2.5: LeNet architecture introduced in LeCun et al. [1998].

increases and the displacement range for each node is reduced. After the final level, the B-spline interpolation outputs the displacement field in the original image size.

The evaluation results of *Deeds* method on different datasets [Heinrich et al., 2013a, 2012b; Xu et al., 2016] presents high accuracy for large deformable motion estimation. However, the computation efficiency of *Deeds* and other classical image registration methods are usually not sufficient for realtime applications. With deep learning techniques, which allows fast inference based on model learning and computation on a GPU, it is now possible to perform image registration of 3D images even under a second. Recently, many image registration methods are developed using deep learning techniques, which includes the methods presented in Chapter 4 and Chapter 5. In the following section, some important deep learning techniques will be introduced, which are particularly relevant to the methods presented in this thesis.

2.2 Image Registration using Deep Learning

Since the introduction of the deep CNN network AlexNet, which has won the ImageNet challenge in 2012 [Krizhevsky et al., 2017], the development of image processing methods using CNNs have increased rapidly. CNN is a class of neural network that can be characterized by their use of shared weights. It is first introduced in LeCun et al. [1998], which applied a CNN model for recognition of handwritten characters (Figure 2.5). In the domain of medical image registration, the first generation of research work using CNNs focused on learning deep features, which substituted similarity measures of conventional image registration methods. These deep features can be combined with iterative registration algorithms and have shown improvements in accuracy. More recently, most researchers use CNNs for a direct prediction of the transformation parameters or complete deformation fields to enable non-iterative inference by feed forward networks (end-to-end learning).

Deep learning approaches in medical image analysis are well introduced in Litjens et al. [2017]; Greenspan et al. [2016] and Ker et al. [2017]. Comprehensive reviews on recent works in deep learning approaches for medical image registration are provided by Fu et al. [2020]; Boveiri et al. [2020]; Haskins et al. [2020]. In Andrade et al. [2018], a review on medical image registration is provided, which also includes an overview on conventional registration as well as deep learning based approaches. In Boveiri et al. [2020], deep learning based approaches are categorized into five generations: deep similarity metrics, supervised end-to-end approaches, deep reinforcement learning (agent-based approaches), unsupervised end-to-end approaches and weakly/semi-supervised end-to-end approaches. According to this, the approach introduced in Chapter 4 falls into the category of weak supervision. Finally, in Chapter 5, we introduce a method that can be seen as a novel direction for an unsupervised registration approach.

In this section, a brief summary of CNNs and the U-Net as well as more details on some example approaches that train CNNs for end-to-end image registration will be given.

2.2.1 Convolutional neural networks (CNNs)

Convolutional Neural Networks (CNNs) have become the most popular deep learning technique for image analysis, due to their powerful and efficient performance in extracting and learning image features from image patches or the whole image [Boveiri et al., 2020]. Typical CNNs consist of multiple sets of convolutional layers, rectified linear units (ReLU), pooling layers, (optionally) batch normalization layers and finally a fully connected layer as the last layer. The output of the last layer before the final fully connected layer is transformed into a vector form. Then, depending on the task, the final output is given as a probability score (for classification) or real values (for regression).

Convolutional layers consist of a small kernel (typically $3 \times 3 \times 3$) and are trained to extract the most significant features from the input via convolution. The *kernel* or *filter* slides over the input image or feature map from the previous layer, and the dot product of the image/feature map and kernel is computed. As a result, a feature map is generated, which consists of low-level features such as edges, dots and lines at the early layers (as shown in Figure 2.6) and in later layers high-level features such as structures. Since the convolution operation is translation invariant, CNNs also inherit this property. In addition, unlike fully connected layers, CNNs are efficient due to their sparsity in connections, i.e. not all input neurons are connected to the next layer and the use of parameter (weight) sharing.

The rectified linear unit (ReLU) is an activation function that ensures non-linearity and prevents vanishing gradient problems during backpropagation. Pooling layers reduce

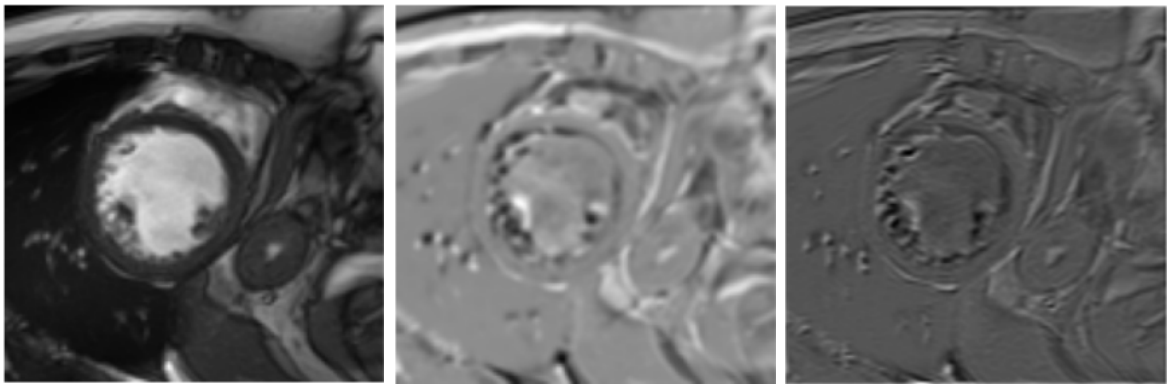


Figure 2.6: Example of filtered images. The original image (left) forwarded to the first convolutional layer from a trained model (right). The image in the middle shows an image filtered using the Sobel filter (for edge detection). In both filtered images, the edges are emphasized. However, these are more distinct in the image on the right.

the dimensionality of the input or the feature maps generated by convolution. With max pooling, the strongest activation over a neighborhood is summarized and with the average pooling, the activation strength over a neighborhood is averaged. The output of this set of layers is a feature map, which is fed into the next set of layers. With the decreasing image size via pooling layers, the receptive field of the image can be increased. Finally, the output feature map is transformed into a vector form for the last fully connected layer, which computes a probability score for classification or values for regression.

As an active object of research, many variations were developed, including CNNs using residual connections, which ensures the unchanged flow of previous features through the layers [Srivastava et al., 2015; He et al., 2016] and feature recycling that concatenate features of different depth directly [Huang et al., 2017]. A comprehensive overview of the evolution of CNN architectures, details of basic CNN components and discussion on applications and challenges of CNNs is provided by Khan et al. [2020]. In this thesis, our goal is to make local decisions, e.g. determining the displacement in each location, rather than performing a global classification, e.g. finding the class of the image. One of the variations of CNN to do this is U-Net, which belongs to the category of fully convolutional architectures and works with an encoder-decoder network model that preserves spatial dimensions.

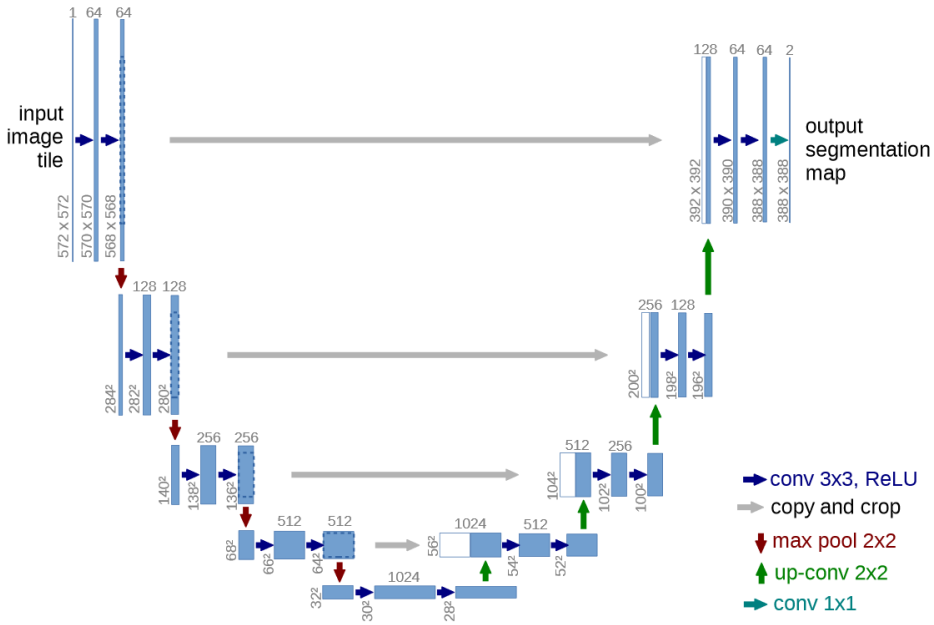


Figure 2.7: U-Net architecture proposed by Ronneberger et al. [2015]. Gray arrows indicate skip connections.

2.2.2 U-Net

In order to achieve a good compromise between the recognition of local object details, understanding of the relationship between larger structures, an encoder-decoder architecture is introduced that produces pixel-accurate outputs and captures spatial context. The encoder-decoder architecture has a symmetrical structure, where in its encoder part the input is contracted, extracting features with high-level abstraction and the feature maps are then expanded back in the decoder part to restore details. With skip connections, which transfer certain feature maps generated in the encoder part directly to the decoder part, this prominent architecture, called U-Net [Ronneberger et al., 2015], can alleviate the loss of detail in contrast to simpler fully convolutional CNNs. The architecture of U-Net is illustrated in Figure 2.7. The U-Net is broadly used for natural or medical image segmentation, and also adapted in many image registration approaches [Hering et al., 2019a; Ha et al., 2020]. Particularly, it enables effective feature learning from a small training dataset, which is a significant advantage for medical applications.

In the following, three examples of prominent state-of-the-art deep learning based image registration methods will be introduced, which are trained with supervision (section 2.2.3), weak supervision (section 2.2.4), and no supervision (section 2.2.5).

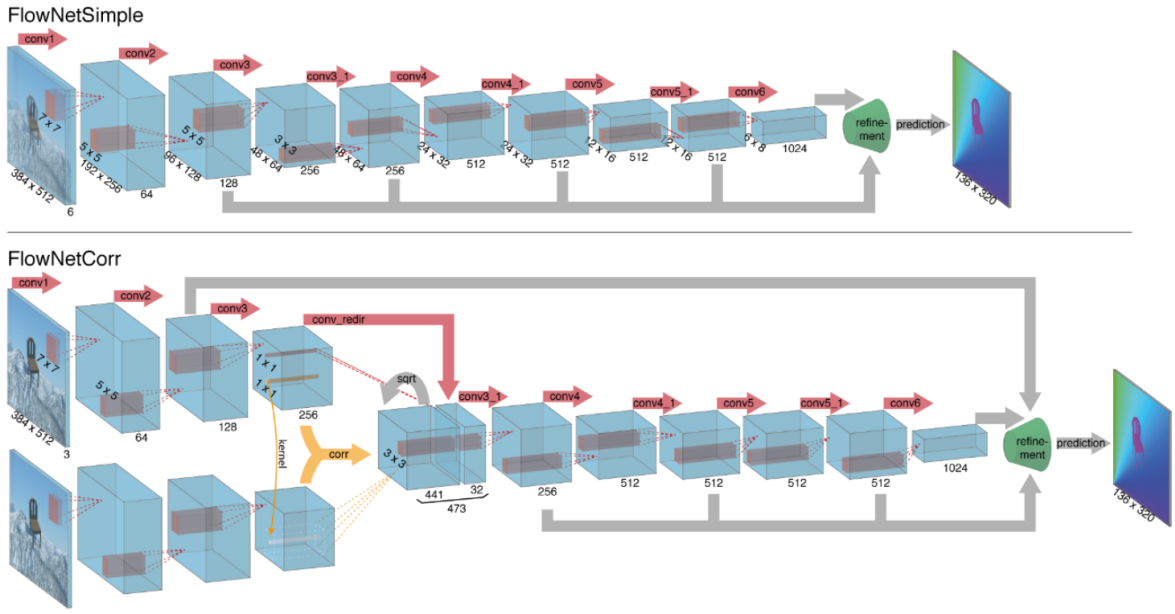


Figure 2.8: Illustration of CNN architecture for FlowNetS and FlowNetC [Dosovitskiy et al., 2015].

2.2.3 FlowNet

FlowNet is an end-to-end supervised deep learning method that aims to estimate optical flow using CNNs [Dosovitskiy et al., 2015]. Optical flow is a method to estimate motion of objects, surfaces, or edges in a scene relative to the observer. An optical flow field can be used for image registration similarly to a deformation field. In Dosovitskiy et al. [2015], two versions of network architecture are introduced: FlowNetSimple (FlowNetS) that only consists of convolutional layers and pooling layers and FlowNetCorr (FlowNetC) that also introduces a non-trainable correlation layer (Figure 2.8). In Ilg et al. [2017], an extended version of FlowNet, FlowNet2.0 is introduced, improving the estimation quality and accuracy by combining the two versions of FlowNet, which is not discussed here.

Both FlowNetS and FlowNetC comprise a contracting part and expanding part. In the contracting part, where convolution and pooling is performed alternately, the network is trained to extract meaningful features of input images that will help estimate optical flow between the input images. In the expanding part, upconvolution and unpooling is performed for refinement of the feature map. The feature maps with the same resolution generated in the contracting part are concatenated before each upconvolution to compensate for the lost details.

FlowNetS has a single path, which takes stacked images as input that go through convolutional and pooling layers together, whereas FlowNetC takes two images separately, processing them using shared weights and then combines them using a correlation layer in the later process. The network is optimized using stochastic gradient descent. Theoretically, when the network is large enough, it is able to learn optical flow under supervision. However, in practice, it cannot be guaranteed that the network trained with local gradient optimization learns the right prediction for large deformations.

FlowNetC solves this problem by constraining the network to first learn meaningful representations of the input images and combining the information at a later stage. Given two images $I_1, I_2 : \mathbb{R}^{H \times W} \mapsto \mathbb{R}$, outputs of the first network streams are feature maps $\mathbf{f}_1, \mathbf{f}_2 : \mathbb{R}^{H \times W} \mapsto \mathbb{R}^C$, where H and W are height and width of the image respectively and C is the number of channels of the features. Given a fixed patch size of $K = 2k + 1$, the correlation c can be computed by convolving the patches from both feature maps as:

$$c(\mathbf{x}_1, \mathbf{x}_2) = \sum_{\mathbf{o} \in [-k,k] \times [-k,k]} < \mathbf{f}_1(\mathbf{x}_1 + \mathbf{o}), \mathbf{f}_2(\mathbf{x}_2 + \mathbf{o}) > . \tag{2.10}$$

Based on the maximal displacement d of the objects between two images, the range of the location $\mathbf{x}_2$ in the second image can be limited, and the correlation $c(\mathbf{x}_1, \mathbf{x}_2)$ can be computed only for the neighboring region. Since the correlation layer does not involve convolution with a filter but convolution between two images, there are no weights to be learned. The output of the correlation layer has the size of $H \times W \times D^2$, where $D = 2d + 1$ is the neighborhood size for which the correlation is computed. The output of the correlation layer is then further processed by the contracting and expanding part of the network, and finally an estimation of optical flow is obtained.

The network is trained using a training loss that computes an endpoint error, i.e. the Euclidean distance between the network output and the ground truth optical flow. For supervised training of a network for image registration, obtaining sufficient ground truth data is essential. In Dosovitskiy et al. [2015], the ground truth optical flows are generated synthetically using different techniques. E.g. using a 3D laser scanner which recorded the motion of moving objects simultaneously while images were taken or using a synthetic dataset with ground truth optical flow (MPI Sintel, Flying Chair). However, for medical images, the only possibility to generate a ground truth deformation field is to use an existing registration algorithm and generate the fields based on given images. Recently proposed supervised medical image registration approaches, such as the method of Rohé et al. [2017], generate a pseudo ground truth by registering the segmented organs to a template image, whereas the approach proposed by Uzunova et al. [2017] uses locality-based shape and appearance models to generate ground truth deformation fields.

The necessity of ground truth data and a sufficiently large number of training samples to train the network is one of the fundamental limitations of supervised learning approaches such as FlowNet and their application to medical image registration. Particularly for deformable transformations, reliable ground truth deformation fields with voxel-wise spatial correspondence are rare and generating such ground truth data is impossible in most cases. Different approaches are proposed to train CNNs using artificial transformations such as random transformation [Eppenhof et al., 2018], traditional registration-generated transformations [Sentker et al., 2018] and model-based registration [Uzunova et al., 2017]. In addition to the ground truth generation problem, there are often only a small number of training samples available, which may lead to over-fitting. Due to these limitations, the most recent research work on medical image registration using deep-learning focuses on either weakly-supervised or unsupervised approaches. For both methods, an example will be introduced in the following.

2.2.4 LabelReg

LabelReg is a weakly-supervised learning approach proposed by Hu et al. [2018b]. Instead of training a CNN network based on ground truth deformation fields under direct supervision, *LabelReg* utilizes a more feasible information for supervision of the network training, i.e. labels (or segmentations) of anatomical structures visible in medical images that are annotated manually by experts.

The proposed CNN network has a U-Net like architecture, where four downsampling steps and four upsampling steps with three skip connections are performed. In each downsampling step, convolution, batch normalization and pooling are performed and the size of the input is reduced to half the input size. The upsampling is performed using transposed convolutions and after the last convolutional layer, no batch normalization or pooling is performed. The network takes stacked image pairs as input and outputs dense deformation fields, which can be used to warp the image or labels. Resamplers such as linear-, cubic- or spline interpolations are used for warping the image or labels. During training, the training loss is computed between the fixed label and the warped moving label, which is transformed using the output of a dense deformation field of the network with an spatial transformer [Jaderberg et al., 2015] that enables differentiable warping. For inference, no labels are required, since the only network input are the stacked images.

Because the corresponding structures of different image pairs might not always be the same for all training samples, each label image used during training is generated as one-hot label and selected randomly during training for each iteration. Furthermore, to avoid the trained network to be over-fitted or under-regularized, the labels are converted

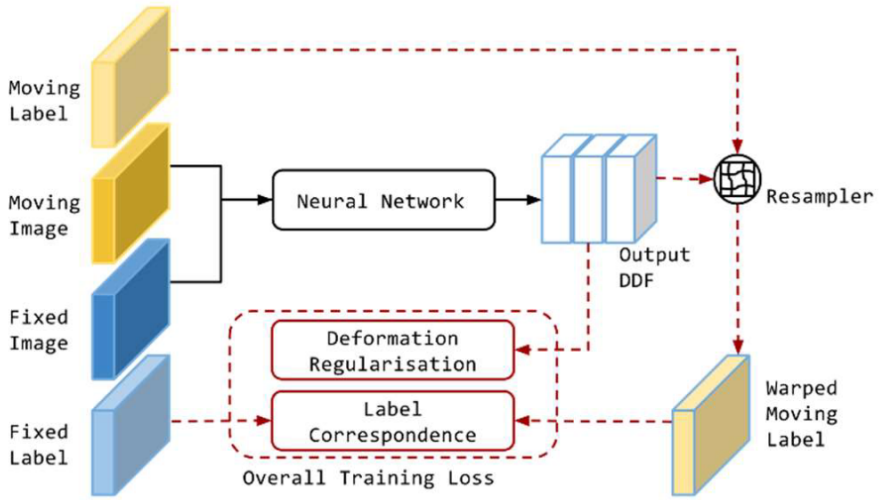


Figure 2.9: Training process of LabelReg [Hu et al., 2018b]. The process illustrated with red dashed line is only performed during training.

into smooth probability maps instead of using binary labels. The foreground of the label image is left unchanged, and only the background is smoothed using the inverse distance transformation. The labels are normalized based on the label with the largest volume.

As shown in Figure 2.9, the network is trained using the typical registration loss formulation, i.e. a combination of similarity and regularization. The second loss term, the regularization loss, ensures the smoothness of the estimated dense deformation field and can be computed using the bending energy or L2-Norm of the displacement gradients. By incorporating the label correspondence loss, the network requires no similarity measures based on image intensities and therefore can be applied regardless of modality difference of input image pairs. However, results can be influenced by the choice of the labels used for loss computation (label-bias).

Similar concepts using segmentations/labels as auxiliary information to estimate dense deformation fields are proposed in the field of computer vision [Cheng et al., 2017; Hur and Roth, 2016; Sevilla-Lara et al., 2016; Tsai et al., 2016]. These approaches jointly estimate segmentations and dense deformation fields in order to deal with the video segmentation problem. For medical images, Qin et al. [2018] proposed a framework for joint estimation of segmentation and motion of a cardiac MR image sequence. They use shared weights for the feature extraction part of the network, which are later fed into two different CNNs for registration and segmentation. Hering et al. [2019a,b] proposed a multi-level deep-learning based registration approach that is trained by combining prior information, such as segmentation, with an energy-based distance metric. The *VoxelMorph* method, which is introduced as an unsupervised approach, has also been

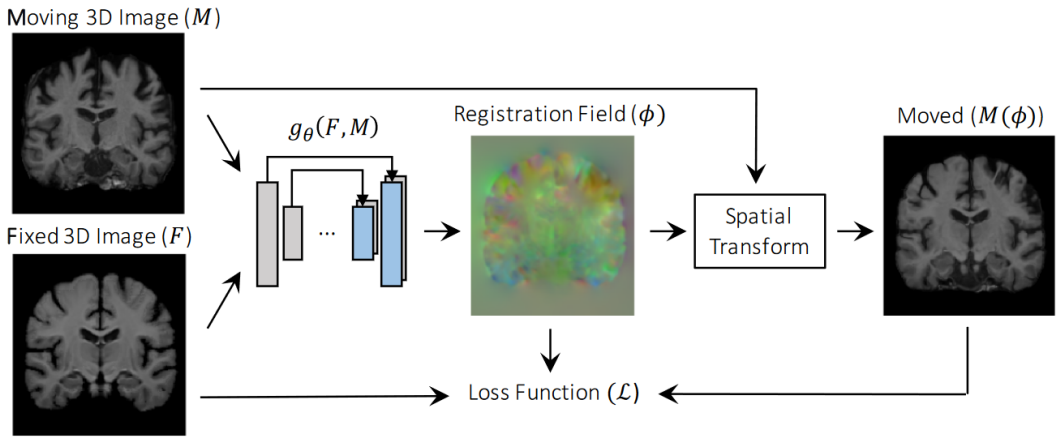


Figure 2.10: VoxelMorph architecture [Balakrishnan et al., 2018].

extended to leverage auxiliary segmentations, using an additional Dice similarity coefficient (DSC) loss function to improve the accuracy of the registration. Our work introduced in chapter 4 also incorporate segmentation as auxiliary information to optimize a CNN network together with other losses, particularly focusing on the use of an explicit segmentation loss. The method is evaluated on MRI cardiac data for registration of end-systolic and end diastolic images and given segmentation of heart structures (ventricles and myocardium), it shows improved results compared to the state-of-the-art unsupervised methods.

2.2.5 VoxelMorph

The last method that is related to the method introduced in this thesis is the *VoxelMorph* approach, which was already briefly mentioned above. Much ongoing research work particularly focuses on unsupervised image registration approaches. *VoxelMorph* is a state-of-the-art end-to-end unsupervised registration approach [Balakrishnan et al., 2019, 2018]. It trains a CNN of U-Net like architecture to learn a global function that maps a pair of input images to a dense deformation field that optimally aligns one image to the other (Figure 2.10).

The most important advantage of the *VoxelMorph* framework is that the network training does not require any ground truth data for supervision. The network is directly optimized by comparing the difference between input images after applying the estimated displacement field on the moving image.

Given training image pairs I_f and I_m, the network parameters θ of a U-Net like network g are updated based on the two loss functions $\mathcal{L}_{sim}$ and $\mathcal{L}_{smooth}$, so that for an unseen image pair, the network can estimate an appropriate displacement field, which

can align the same structures in both images. Similar to the FlowNet and LabelReg, the network consists of a contracting (or encoding) part and an expanding (or decoding) part. In the encoding part of the network, the input images are downsampled by a stride of 2 and the receptive field of the network increases, so that the size of the receptive field is at least as large as the maximum expected displacement between the fixed and moving images. In the decoding stage, the features learned in the encoding phase are concatenated with the outputs of the decoding layer using skip connections.

The loss terms that are used to update the network parameters are similar to the energy function of classical registration methods. $\mathcal{L}_{sim}$ can be computed using any similarity measure that penalizes the difference between two images. In the original *VoxelMorph* work [Balakrishnan et al., 2018], the authors use a negative local cross correlation function that is robust to the variations in image intensity. To enable end-to-end network training, the differentiable warping step using a spatial transformer [Jaderberg et al., 2015] is performed on the moving image. As for the regularization term $\mathcal{L}_{reg}$, which penalizes local spatial variation and encourages a smooth deformation field, a diffusion regularization on spatial gradients is used.

Other similar approaches are proposed by de Vos et al. [2019]; Dalca et al. [2019]; Krebs et al. [2018]. They utilize conventional similarity metrics and spatial transformers to train end-to-end networks in an unsupervised manner. Dalca et al. [2019] extends the *VoxelMorph* approach using a probabilistic generative model, which ensures diffeomorphic deformation fields. Krebs et al. [2018] train a convolutional variational autoencoder (CVAE), also in a probabilistic and generative fashion, to obtain a diffeomorphic deformation field. Multiple ConvNets are stacked into a larger network to train a multi-stage image registration framework for affine and deformable transformation in de Vos et al. [2019].

As a result of image registration, we usually obtain a dense deformation field or transformation parameters. However, since the preparation of ground truth data for such results are difficult, as mentioned in the previous section, the evaluation of the developed method cannot be performed directly by comparing the ground truth data and the registration results. Therefore, different metrics, which are based on segmentation or landmarks, are usually used for quantitative and/or qualitative evaluation to measure the accuracy and plausibility of the registration results.

2.3 Evaluation Metrics for Image Registration

In this section, a short explanation of the evaluation metrics used for quantitative or qualitative measurement of registration performance will be given, focusing on the metrics that are used in this thesis.

To evaluate the performance of a registration method, quantitative evaluation metrics such as Dice coefficient, mean contour distance, and target registration error (TRE) can be computed. These metrics measure the accuracy of the estimated deformation field, usually by computing quantitative values by comparing the annotations of two images obtained manually by medical experts.

2.3.1 Dice coefficient

Calculating the Dice coefficient is a segmentation based evaluation method, which is equivalent to the F1 score. To compute a Dice coefficient, ground truth labels for both images of image pairs are required. As a ground truth label, important image objects such as anatomical structures in the image are usually segmented manually. Given the label images $A, B \in \Omega$ and the estimated transformation $\mathcal{T}$, where $\Omega = \{0, 1, ..., L\}$ with L being the number of segmented image objects, the Dice coefficient can be computed as:

$$Dice_l = \frac{2\sum_l |\mathcal{T}(A_l)B_l|}{\sum_l |\mathcal{T}(A_l)| + \sum_l |B_l|} \tag{2.11}$$

where $\mathcal{T}(A_l)$ and B_l are the pixels of label images $\mathcal{T}(A)$ and B with the label l respectively. The resulting value range from 0 to 1, with 1 indicating the perfect match. The mean Dice coefficient can be obtained by averaging the Dice coefficients for all labels.

2.3.2 Mean contour distance

Another segmentation based evaluation method is the mean contour distance, also referred to as average surface distance (ASD). The contour of an image object can be generated from the segmentation image or from landmarks set around the objects. The contour of a structure is defined as a group of pixels on the border of the structure. Given the set of contour distances S (see below) of each structure, the mean contour distance can be computed as:

$$D_{contour} = \frac{\bar{d}(S(\mathcal{T}(A)), S(B)) + \bar{d}(S(B), S(\mathcal{T}(A)))}{2} \tag{2.12}$$

where $S(A)$ denotes the contour of the label image A and $\bar{d}(X,Y)$ for some contour X and Y is defined as:

$$\bar{d}(X,Y) = \frac{1}{|X|} \sum_{x \in X} \min_{y \in Y}(x,y). \tag{2.13}$$

where $|X|$ is the total number of pixels in contour X. The mean contour distance is computed in pixel/voxel units, however, can be easily converted into mm, when pixel/voxel spacing is given.

2.3.3 Target registration error (TRE)

The TRE is a landmark-based evaluation method, which can be used when no contour/segmentation annotation for specific structure is available. Landmarks can be set on geometrically prominent anatomical points such as bifurcations, a tip of an organ, or the end point of bones. The TRE is computed between the set of N landmarks $L^A, L^B \in \{(x,y)_{j=[0,...,N]} | (0,0), ..., (H,W)\}$ of image A and B, which are set on the corresponding points of the image pairs.

$$mTRE = \frac{1}{N} \sum_{j=1}^{N} ||L_j^A - L_j^B||_2. \tag{2.14}$$

where L_j^A and L_j^B are the j-th landmark of image A and B, respectively.

2.3.4 Jacobian determinant

In addition to the alignment accuracy, the quality of the estimated transformations can be evaluated using Jacobian determinants, which is crucial for subsequent visual assessment of deformations for longitudinal analysis. The Jacobian determinant (also *Jacobian*) provides a quantitative evaluation of the topology of the deformation field.

For each pixel (x,y), $D(x,y)$ is the estimated displacement vector.

$$Jacobian = det \left[I + \begin{pmatrix} \frac{\delta D_x(x,y)}{\delta x} & \frac{\delta D_x(x,y)}{\delta y} \\ \frac{\delta D_y(x,y)}{\delta x} & \frac{\delta D_y(x,y)}{\delta y} \end{pmatrix} \right] \tag{2.15}$$

defines the Jacobian for the 2D case, where I is the identity matrix, $D_x(x,y)$ and $D_y(x,y)$ are the x and y component of $D(x,y)$ respectively. For the 3D case with each voxel (x,y,z) and displacement vector $D(x,y,z)$, the Jacobian is defined as:

$$Jacobian = det \left[I + \begin{pmatrix} \frac{\delta D_x(x,y,z)}{\delta x} & \frac{\delta D_x(x,y,z)}{\delta y} & \frac{\delta D_x(x,y,z)}{\delta z} \\ \frac{\delta D_y(x,y,z)}{\delta x} & \frac{\delta D_y(x,y,z)}{\delta y} & \frac{\delta D_y(x,y,z)}{\delta z} \\ \frac{\delta D_z(x,y,z)}{\delta x} & \frac{\delta D_z(x,y,z)}{\delta y} & \frac{\delta D_z(x,y,z)}{\delta z} \end{pmatrix} \right] \tag{2.16}$$

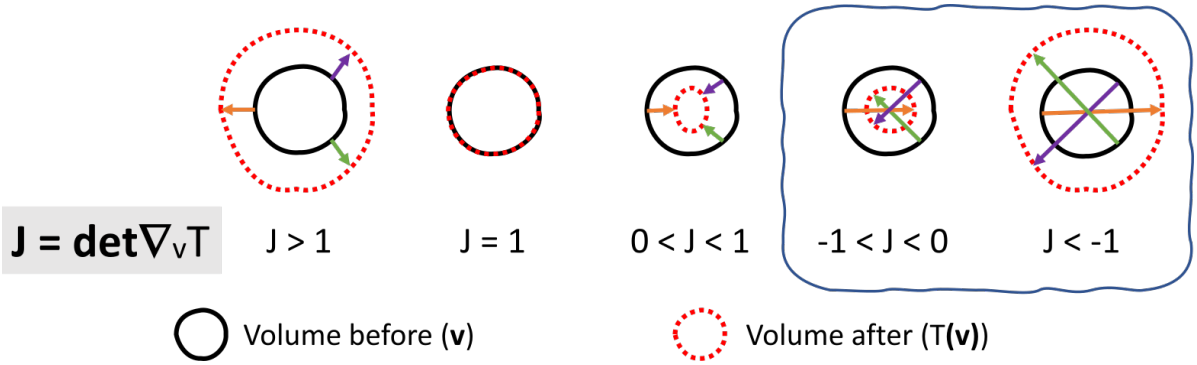


Figure 2.11: Schematic illustration of deformation characteristics indicated by the Jacobian values. The black circle is the original volume and the circle drawn with red dotted line is the volume after the deformation. Two cases illustrated in the box are the cases where the deformation is considered to be implausible for medical image registration.

where I is the identity matrix and $D_x(x, y, z)$, $D_y(x, y, z)$, and $D_z(x, y, z)$ are the x-, y- and z-component of $D(x, y, z)$ respectively.

A Jacobian in each pixel provides the characteristic of the deformation, i.e., a Jacobian of 1 denotes no change, > 1 indicates expansion, $0 - 1$ shrinkage, and a Jacobian < 0 indicates singularity as illustrated in Figure 2.11. If the standard deviation of the Jacobian has a small value, the transformation is smooth. The number of negative Jacobian determinants indicates the number of singularity points. When there should be no occlusion of the structures, which is the case in most medical image registration tasks, the smaller mean number of pixels with a negative Jacobian indicates a plausible deformation.

The choice of evaluation metrics depend on the specific task/problem, for which the registration method is applied. In the following section, we will give some background information on the medical applications considered in this thesis.

2.4 Medical Background

Image registration is required in various stages of medical procedure for different purposes. In this work, the focus is on the application of image registration for image-guided interventions such as image-guided radiation therapy, image-guided High Intensity Focused Ultrasound (HIFU) and image-guided surgery. For such treatments or procedures, image registration is required in each step from the planning to the post assessment in order to improve the treatment result and to minimize the risks and side effects such

as unwanted damages on healthy tissues. In this work, we focus on the intraoperative image registration with the preoperative images for improvement of the image-guidance. For such an application, the speed of registration is of significant importance in addition to the accuracy, in order to enable a smooth workflow and to prevent time delays of the clinical procedures. Particularly, to give medical practitioners the ability to adapt their plans based on the guidance images, it is important to be able to register images in realtime.

In the following, a short introduction to image-guided interventions will be presented (section 2.4.1) as well as some details on the specific applications such as MRI-guided radiation therapy (section 2.4.1.1) and ultrasound-guided brain surgery (section 2.4.1.2).

2.4.1 Image-guided interventions

Image-guided interventions are applied in clinical practice e.g. during surgery, radiation therapy, or biopsy. It is performed with a computer-based system to aid visualization of the region of interests (ROIs) across various imaging modalities. The use of image-guidance enables minimally invasive procedures by visualizing anatomical structures without opening-up a patient's body. The typical procedure includes preoperative image acquisition of the patient, usually using 3D tomographic images with a high spatial resolution, which take longer time for acquisition but provide more details than the guidance images. Although intraoperative guidance images acquired during the procedure have a lower spatial resolution, they are taken much faster to provide realtime information on the changes.

Before an actual operation is performed on the patient, preoperative images are taken to plan an appropriate procedure. The images acquired in this phase provide detailed anatomical and/or pathological information on the patient and the region to be treated. In the actual procedure, several steps have to be done to set the image-guidance system ready. First, patient and surgical instruments have to be localized. In radiation therapy, for instance, the patient table is adjusted to the position as in preoperative image acquisition with the help of reference marks [Saenz et al., 2018]. For the localization of surgery instruments, different types of tracking systems are available including optical videometric, infrared, and electromagnetic systems. After the localization of the patient and/or surgical instruments, the patient's anatomy is registered to the preoperative images. In this step, an accurate, robust, and ideally fast image registration algorithm is required to spatially align the intraoperative images with the planning images. The most widely used image registration approaches in clinical image-guided systems is still the rigid registration approach that works sufficiently well for rigid bony structures (e.g. spine or brain surrounded by rigid cranium). However, it might not be the best solution

for preoperative of the regions such as thorax and abdomen, where the deformations are elastic. In this regard, an increasing amount of research has been performed on nonrigid deformable image registration approaches in the last decade and some commercially available image-guided systems now provide a software with a deformable image registration algorithm [Mittauer et al., 2018; Winkel et al., 2019]. Once the patient's anatomy is registered to preoperative images, surgical instruments are positioned and displayed on the guidance image relative to the patient's anatomy. In radiation therapy, for instance, the target volumes are delineated based on the registered guidance images. Throughout the procedure, the realtime guidance images are acquired and used to adapt the planned procedure according to changes in patient anatomy and enable the online adaptation of the treatment plan.

Image-guided interventions have a wide range of application possibilities for many different regions and organs in the body. In the following, we introduce two examples of image-guided interventions in detail, which includes MRI-guided radiation therapy and ultrasound-guided brain surgery. Our proposed image registration approaches introduced in Chapter 3 to Chapter 5 are evaluated on publicly available datasets related to these interventions with different scenarios.

For MRI-guided radiation therapy, the goal was to register intraoperative guidance images taken under free-breathing to the preoperative planning images to enable realtime monitoring of patient's (respiratory) motion (intra-fractional movement) in the thorax and abdomen region. Particularly for organs located in thorax and abdomen, respiratory motion is the most important factor that influences the accuracy of the treatment and should be monitored continuously during treatment in order to adapt the plan accordingly and to minimize radiation exposure to healthy tissues and organs-at-risk.

For image-guided neurosurgery, we evaluate our method for the registration of 3D preoperative MRI data to 3D intraoperative ultrasound data. Although the brain can be fixed during a procedure, brain anatomy can still shift particularly after craniotomy, opening of the skull, due to the change in pressure and during the procedure due to lesion resection, bleeding, and fluid drainage. Realtime monitoring and adaptation of surgery plan according to such changes can improve the quality of the surgery as well as lead to a better prognosis. For this purpose, an accurate and robust realtime registration algorithm is essential.

2.4.1.1 MRI-guided radiation therapy

In external beam radiation therapy, the aim is to destroy a tumor by delivering a high dose of radiation to it. In the meantime, irradiation of healthy tissues around the tumor and nearby organs-at-risk (OAR) should be minimized. Challenges in radiation therapy

Figure 2.12: MRI-guided radiotherapy system *MRIdian* by ViewRay. (Image from [Klüter, 2019])

come from anatomical changes of the patient between each treatment (inter-fraction) and/or between planning and treatment, as well as during treatment (intra-fraction).

Respiratory motion is one of the important types of intra-fractional motion that has to be accounted for during radiation therapy when treating a tumor in the thoracic and abdominal region. Although the most commonly used image registration methods perform rigid image registration, for compensation of respiratory motion deformable image registration is required for accuracy. Currently, different motion management strategies are used including breath-hold, respiratory gating, and motion tracking to account for the tumor movement due to respiratory motion. Both breath-hold and respiratory gating strategies control radiation beams by turning it on when the tumor is at the predefined location. For the breath-hold approach, the patient is usually trained before the treatment and coached using audio- or visual guidance to hold their breath at a certain phase of the respiratory cycle. When using respiratory gating, the patient can breathe freely and a beam management is performed based on guidance images. The guidance images should be registered with the planning image to be able to re-optimize the current treatment plan.

The best motion management can be made via motion tracking, which aims to account for tumor motion continuously and dynamically, thereby reducing the total treatment time and improve the accuracy of dose delivery. To track tumor motion, guidance images are employed using internal imaging modalities such as US, X-ray, or external imaging

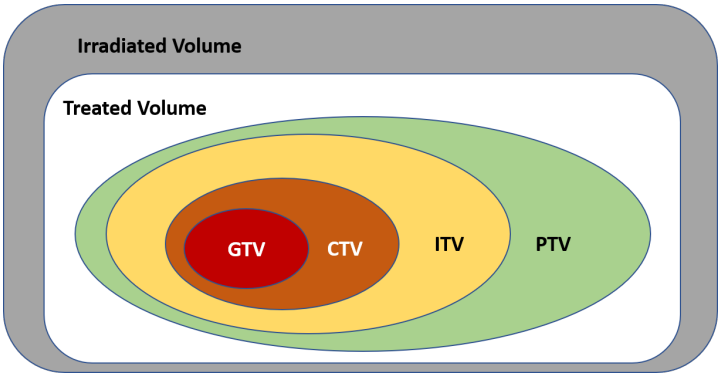


Figure 2.13: Illustration of target volume definitions from Landberg et al. [2016].

e.g. infrared, or video cameras combined with markers to be tracked that have been implanted in the patient's body. However, performing non-invasive motion tracking solely based on these guidance systems is only partially feasible due to problems such as low image quality, limited field-of-view, and the risk caused by extra irradiation.

Recently, image-guided external beam radiation therapy systems with integrated MRI for guidance have been developed by different groups and are made available for clinical use (e.g. *MRIdian* by ViewRay shown in Figure 2.12 and *Unity* from Elekta). Although the acquisition time of MRI is longer than CT or ultrasound images, it has excellent soft-tissue contrast and no risk of extra irradiation during image acquisition. These properties enable MRI-guided radiotherapy systems to directly visualize and monitor tumor and OAR motion without extra markers and the damages due to additional radiation. More details on MRI-guided radiation therapy systems, their current development, their advantages, and remaining challenges are explained in Otazo et al. [2020].

In typical radiotherapy planning, the target volume is delineated including margin volumes as defined in International Commission on Radiation Units and Measurements (ICRU) [Landberg et al., 2016] on the planning images. A schematic illustration of the target volume definitions is depicted in Figure 2.13. According to the volume definitions by Landberg et al. [2016], the target volume consists of the gross tumor volume (GTV), the clinical target volume (CTV), the internal tumor volume (ITV), and the planning target volume (PTV). GTV is the visible (tumor) volume that can be distinguished with eyes (by an experienced radiologist), whereas CTV includes margins for possible microscopic malignant tissues. ITV includes a margin around the GTV, CTV, and OAR to account for variations in volume, shape and position caused by the patient motion. PTV can be delineated by accounting for the uncertainties and reproducibility of system settings and the daily set-up. This adds additional margins to the ITV. With an advanced motion tracking method integrated into the online adaptation workflow, ITV margins can be reduced, resulting in smaller PTV and sparing more healthy tissues.

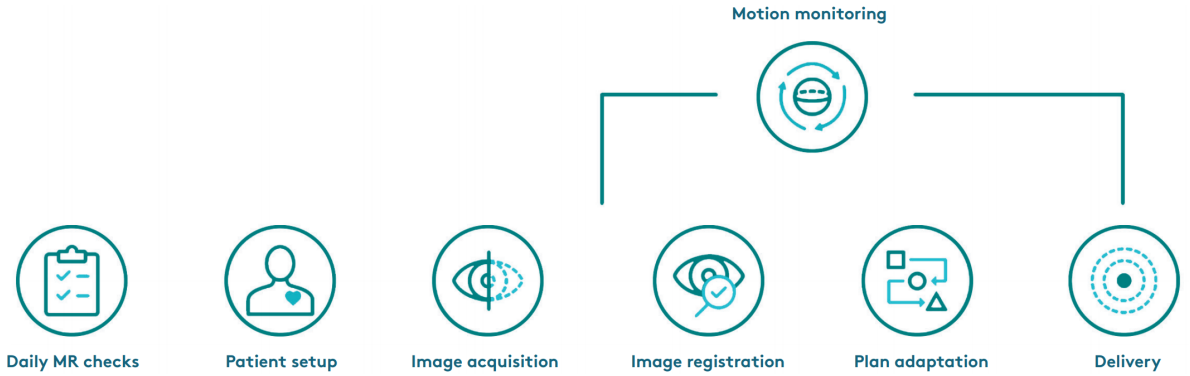


Figure 2.14: An overview of online adaptive radiation therapy workflow. (Image source: [Brwon and de Vries, 2018])

An example online adaptive radiotherapy workflow from *Unity*, Elekta is shown in Figure 2.14. When a deformable image registration can be performed in realtime, the treatment plan can also be adapted on the fly. However, traditional deformable image registration approaches usually take a long time to register an image pair and is inapplicable to this type of task.

2.4.1.2 Ultrasound-guided brain surgery

In neurosurgery, image-guidance is primarily used to locate an intracranial lesion for biopsy or resection. For diagnosis, MRI images are usually taken to determine information on the tumor (e.g. size and position) and intra-path anatomy, i.e. the anatomical structures between the tumor and the surface, for surgical procedures. However, these images cannot be used directly for surgical guidance, because of possible brain shift after craniotomy due to drained fluids, bleeding, decompression of cyst, and further complications. Therefore, guidance images should be registered onto the planning images before starting the resection to adopt the plan according to the new locations of anatomical structures. A neuronavigation system is used to perform this registration before starting the operation, which registers the pre- and intraoperative imaging data.

For both the neuronavigation system and image-guidance during treatment, different imaging modalities can be used, such as intraoperative MRI or intraoperative US [Hu et al., 2018a]. Using intraoperative MRI, the outcome of brain surgery can be improved significantly [Hlavac et al., 2017]. However, there are some drawbacks such as longer duration of the surgery due to the acquisition and computation time for MRI image, the need for special surgery instruments, and expensive costs. On the other hand, US images can be acquired in realtime without a pause of the procedure, the reconstruction

of the image can be made quickly and repeatedly. For this reason, a robust multi-modal image registration is important, since it enables combination of both modalities.

In this chapter, essential background information including the general image registration process and its components, image registration using deep learning techniques, some commonly used evaluation metrics, and a brief introduction to some relevant medical applications to better understand this thesis have been given. In the following three chapters, the registration methods developed within the scope of image-guided interventions will be presented. Starting with the classical image registration method, which aims to enable realtime estimation of respiratory motion for image-guided interventions, the following two chapters will present deep learning based approaches, which require no ground truth displacements.

Chapter 3

Model-based Sparse-to-Dense Deformable Image Registration

In this chapter, a model-based deformable image registration framework for image-guided interventions published in [Ha et al., 2018] and [Wilms et al., 2016] will be presented. The proposed framework is evaluated on the scenario, where prior knowledge on deformation domain can be extracted from the existing data. Wilms et al. [2016] and Wilms [2018] present a novel coupled convex optimization algorithm for integration of this prior knowledge into the deformable image registration framework to enable plausible estimation of deformation field. As an extension of above-mentioned works, the work presented in this chapter especially focuses on speeding up of the computation time to meet realtime requirement, which is essential for online motion compensation during treatment. With GPU programming and implementation of an efficient block-matching algorithm, a significant improvement in computation time is achieved.

3.1 Introduction

Respiratory motion is one of the important factors that has to be considered during the course of radiation therapy using a linear accelerator (Linac) or high intensity focused ultrasound (HIFU), especially when they are applied on thorax or abdomen regions. To account for the respiratory motion during radiation therapy, different strategies are used such as breath-holding, respiratory gating and motion tracking. While breath-holding and respiratory gating is more common in clinical use currently, they are only a suboptimal solution since the treatment time is increased with these strategies. Motion tracking strategies aim to provide information on the target position during the treatment continuously under free-breathing condition and can therefore reduce the treatment time. Most of the currently clinically available motion tracking systems rely on optical markers

to track the target position [Korreman, 2015]. To use optimal markers for target tracking establishment, the correspondence between the motion and position of the optical markers and the target is required. Poor or erroneous correlation between these two can lead to a decrease in tracking accuracy. Moreover, in some cases where the markers are embedded near the target in the patient body, it might cause an extra problem after the treatment such as inflammation, pain and bleeding [Gill et al., 2012; Loh et al., 2015].

The best solution to avoid above-mentioned problems is to solely rely on the guidance images obtained during the treatment to monitor and track realtime movement of the target and organs at risk. Deformable registration of the images from the different time frames can provide information on the target/organ movements. There are algorithms that have reached intra-observer accuracy on CT data for offline respiratory motion estimation [Ruhaak et al., 2017]. However, these algorithms are computationally complex and cannot be employed for intra-interventional motion estimation that requires the image registration to be done within a second. Moreover, motion estimation based on MRI and US images constitutes additional challenges caused by (1) a high level of image noise and spatially varying contrast, (2) a different field of view or image dimension between the pre-treatment reference (or fixed) image and the intra-interventional template (or moving) images. These two problems are usually not addressed by the classical algorithms developed for volumetric CT images. The first challenge can be solved by employing contrast-invariant feature descriptors [Heinrich et al., 2013b]. The second challenge is more complex to deal with. Rapid change in the field of view in US-guided interventions due to the changes in probe position and organ movement should be considered along with the target motion. For MRI-guided interventions, the guidance images are mostly obtained in 2D slices to enable sufficient temporal resolutions (3-8 Hz) [Stemkens et al., 2016], whereas the pre-treatment images are taken in 3D. To obtain dense 3D motion estimation of the target and organs at risk required for image-guided dose delivery and replanning, a way to deal with these incomplete or sparse information from guidance images should be considered.

3.1.1 Related works

For realtime motion tracking, the most common approach is to track a single or several locations that are relevant to the target using template or block-matching algorithms [Cerviño et al., 2011; De Luca et al., 2012; Bjerre et al., 2013; De Luca et al., 2013; Luca et al., 2015; Brix et al., 2014; Paganelli et al., 2015; Banerjee et al., 2015; Shepard et al., ress]. These approaches can track the target location fast and accurately, however cannot provide a dense motion estimation for the whole image field of view for information on the movement of important organs at risk. Template- or block-matching algorithms

find correspondences between two images solely based on the local image information, and therefore liable to uncertainties that might be present in the matching process. For global spatial regularity and/or temporal smoothness, an appropriate regularization scheme should be combined with these algorithms.

There have been many researches on estimation or interpolation of dense motion fields based on the sparse correspondences obtained using template- or block-matching algorithms. Approaches such as Thin-Plate-Splines [Lee and Krupa, 2011] or piece-wise affine warps [Royer et al., 2017] rely on general, unspecific interpolation, whereas some other approaches employ statistical motion models [McClelland et al., 2013]. The later approaches can incorporate the patient-specific information into the regularization step by statistical motion model generated from the (patient-specific) data obtained in pre-treatment phase [King et al., 2012; Klinder and Lorenz, 2012; Boye et al., 2013; Preiswerk et al., 2014; Stemkens et al., 2016]. These approaches usually have a better reconstruction capability than unspecific interpolation approaches, accurately reconstructing local details even in the areas with very sparse correspondences. In addition, a patient-specific motion model can also serve as a motion prior during the fitting process [de Senneville et al., 2012; King et al., 2012; Stemkens et al., 2016].

Among classical dense registration approaches Demons algorithm [Somphone et al., 2014], advanced optical flow techniques [de Senneville et al., 2015; Seregni et al., ress; Zachiu et al., 2015] or variational image registration approaches [König et al., 2014] enable fast, accurate and robust dense motion estimation using classical regularizers. However, these approaches perform gradient descent-based optimization, which is prone to local minima. Moreover, they cannot be applied for motion estimation when the images have different spatial dimensions. The best performing algorithm on the CLUST 2014 and CLUST 2015 3D ultrasound tracking challenge data sets [De Luca et al., 2015] is the approach from Royer et al. [2017]. Their approach combines intensity-based mesh model fitting with a mechanical regularization and uses gradient descent-based optimization to minimize the cost function and therefore is also prone to local minima.

3.2 Proposed Method

An Overview of the proposed motion compensation framework is shown in Figure 3.1. The section number is given for important components, where the detailed explanations are provided.

Previous to actual treatment, the patient's image of the treatment region is acquired for treatment planning. In this work, we assume that the images acquired in the pre-treatment phase are from the same imaging modality as the images acquired during the treatment (mono-modal registration). For target delineation, dose calculation and

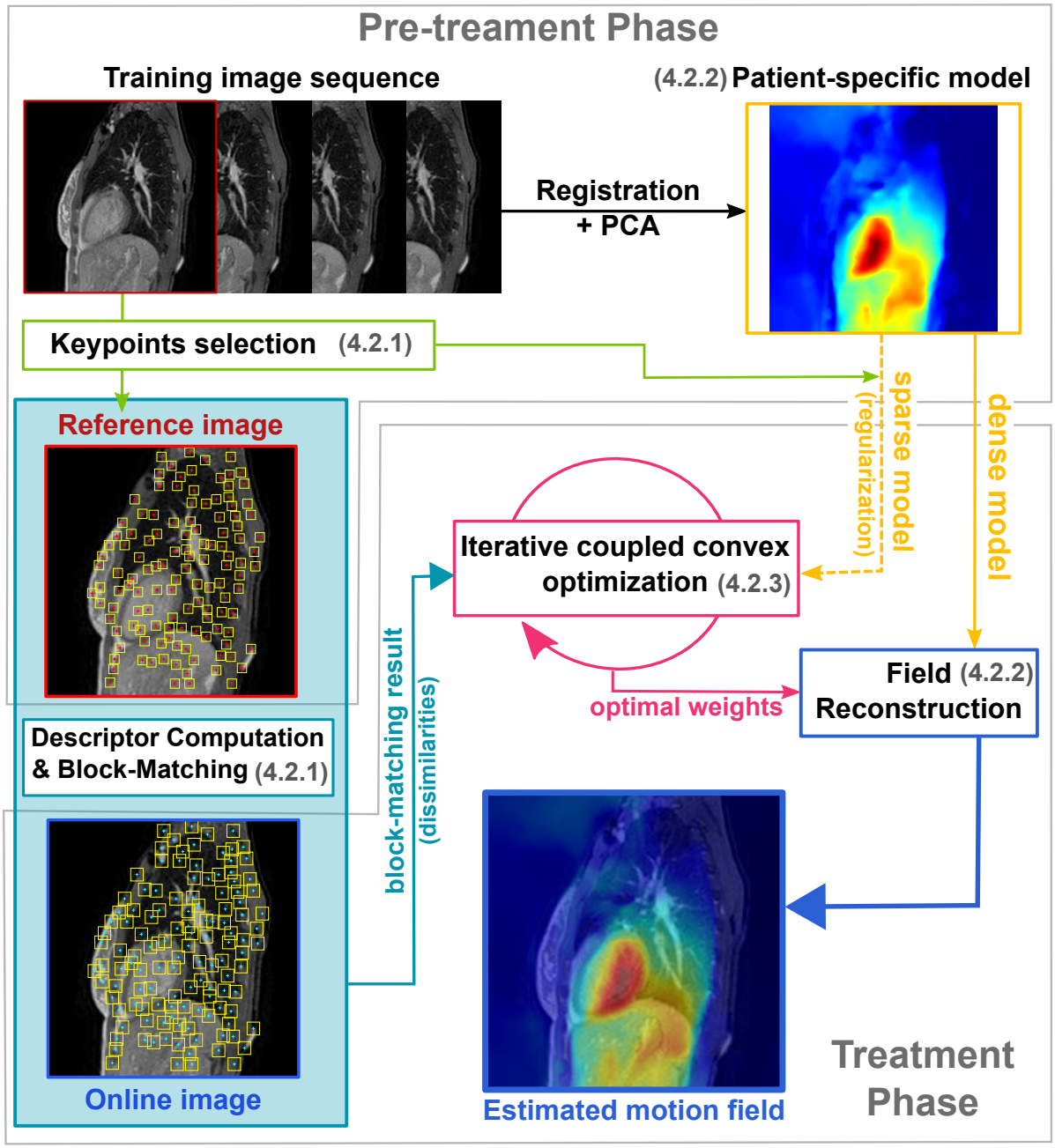


Figure 3.1: An illustration of realtime model-based deformable registration framework. For each component, the section number with detailed explanation is given. (Image source: [Ha et al., 2018])

treatment planning, n-dimensional sequence images of free-breathing patient is acquired for T time frames. The image sequence $\{I_{j^*}\}_{j^* \in \{1,..,T\}}$ is used to generate a patient-specific motion model in our framework and are the training dataset. In addition, a reference image $I_R \in \{I_{j^*}\} : \Omega \to \mathbb{R}$ is selected. The image domain Ω is the region of

interest or the image field of view, for which the motion has to be tracked during the intervention (3.2.2).

During the actual treatment of an image-guided intervention (treatment phase), intra-interventional images (moving images) $I_{M,t} : \Omega' \rightarrow \mathbb{R}$ are acquired at different time points t in realtime. The domain Ω' of intra-interventional images is usually smaller than the domain of the reference image $\Omega' \subseteq \Omega$, having fewer slices and/or lower spatial resolution. Direct determination of a dense transformation $\phi_t = Id + \mathbf{u}_t : \Omega \rightarrow \Omega$ with a dense displacement field $\mathbf{u}_t : \Omega \rightarrow \mathbb{R}^n$ is therefore not possible. To cope with this problem and to satisfy realtime requirements, only N sparse set of keypoints in the reference image are selected $\Omega'_N = \mathbf{x}_1, .., \mathbf{x}_N$ and tracked during the treatment (3.2.1).

Given N sparse set of keypoints set on the reference image, the goal of our framework is to first determine an optimal sparse displacement field $\tilde{\mathbf{u}}_t : \Omega'_N \rightarrow \mathbb{R}^n$ at the time point t that minimizes the following cost function,

$$E(\tilde{\mathbf{u}}_\mathbf{t}) = \sum_{\Omega_N} \mathcal{D}(I_R, I_{M,t}, \tilde{\mathbf{u}}_t) + \alpha \mathcal{R}(\tilde{\mathbf{u}}_t) \tag{3.1}$$

where $\mathcal{D}$ is a point-wise dissimilarity measure between $I_R(\mathbf{x}_i)$ and $I_{M,t}(\mathbf{x}_i + \tilde{\mathbf{u}}_t)$ around the keypoint $\mathbf{x}_i$. $\mathcal{R}$ is a regularization term controlled by the weight parameter α. The regularization term $\mathcal{R}$ regularizes and penalizes deviations of the sparse displacement field $\tilde{\mathbf{u}}_t$ from plausible deformations, and our patient-specific motion model is applied for this purpose. The patient-specific motion model is also used for final reconstruction of the dense deformation field $\mathbf{u}_t$ (3.2.2).

The joint cost function in Equation (3.1) is a non-linear due to our dissimilarity function and cannot be minimized in one step. We propose a iterative scheme to minimize the cost function, called a coupled convex optimization. Using a coupled convex optimization approach, an optimal sparse displacement field can be determined efficiently by alternating optimization over local image dissimilarity distribution and global model-based regularization via motion model (3.2.3).

3.2.1 Keypoints selection and similarity-driven block-matching

To extract keypoints Ω'_N within the reference image selected from the image sequence of the pre-treatment phase, we use the Harris corner detector. However, any automatic landmark detection algorithm that returns distinguishable points can be used. To assure a good distribution of the keypoints, a non-maximum suppression on the result of the Harris corner detector is employed. A non-maximum suppression algorithm selects only the keypoints with maximum values within a certain radius, and the number of the final extracted keypoints depend on this radius. Example images depicting the response of

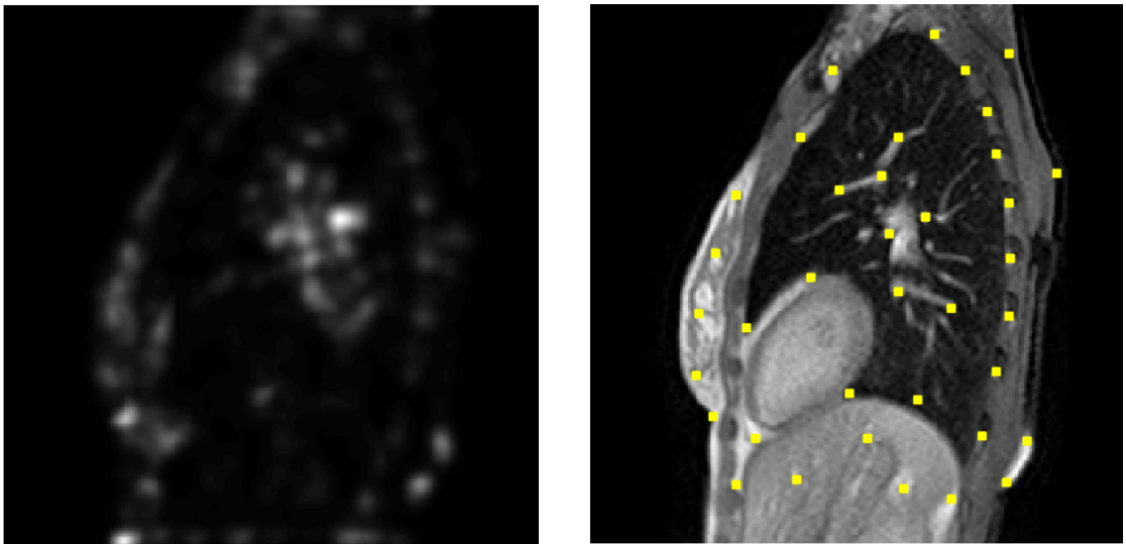


Figure 3.2: Example images depicting the response of the Harris corner detector (left) and the selected sparse keypoints (right).

the Harris corner detector and the selected keypoints are shown in Figure 3.2. In this work, a suitable radius was chosen empirically.

For each keypoint $\mathbf{x}_i$, the dissimilarity cost $\mathcal{D}$ is computed using the block-matching algorithm. The block-matching algorithm compares features of an image block with a certain size that contains the pixels around the considered keypoint $\mathbf{x}_i$ in the reference image to the blocks around the points $\{\mathbf{d}_i\}_t$ within a search region of the moving image. A cost map can be generated for each keypoint, which has the size of the search region and contains the dissimilarity values for each displacement. The initial sparse motion field is determined from the cost map by selecting $\mathbf{d}_{i,t}$ with the minimal cost, and the cost map is stored to be used for iterative optimization (3.2.3).

The result of the block-matching is influenced by the choice of the features used for comparison, and directly using image intensities might lead to suboptimal results due to the presence of the noise and the intensity variation. Therefore, we use image features that are invariant to intensity variation or noise and describe the context of within the neighborhood: the self-similarity context (SSC) descriptors [Heinrich et al., 2013b]. The self-similarity $\mathcal{S}(I, \mathbf{x}, \mathbf{y})$ for a patch centered at the point $\mathbf{x}$ is defined as:

$$\mathcal{S}(I, \mathbf{x}, \mathbf{y}) = exp(-\frac{SSD(\mathbf{x}, \mathbf{y})}{\sigma^2}) \quad \mathbf{x}, \mathbf{y} \in \mathcal{N} \tag{3.2}$$

where $\mathbf{y}$ is the center of a patch within the neighborhood $\mathcal{N}$, σ is a local or global noise estimate and SSD is the sum of squared differences. The SSC descriptors are computed between the six neighbor patches (for three-dimensional image) within the

same distance in pairwise manner, and the considered patch at the center is not included in the calculation.

The initial sparse motion field $\tilde{\mathbf{u}}_t$ determined using the block-matching algorithm will have displacement vectors that are inconsistent with the global motion, since the block-matching algorithm finds the best correspondence for each keypoint without considering the consistency with the other keypoints. To obtain a spare motion field that is more plausible, which can describe both the local and global motion, a model-based regularization is incorporated using the patient-specific motion model generated based on the training images.

3.2.2 Patient-specific motion model

To build a patient-specific respiratory motion model, a non-linear registration algorithm is required. Using the registration algorithm, the training images $\{I_j\}_{j\in\{1,..,T\}}$ are registered to the reference image I_R and dense motion fields for all training images can be obtained. The motion model can be then built from the estimated dense motion fields by performing a PCA. Any non-linear registration algorithm can be used to obtain dense motion fields, however, it should be considered that the accuracy of the chosen algorithm has a great influence on the final result. In this work, we use the publicly available *deeds* algorithm [Heinrich et al., 2013a] that is shown to have high accuracy for respiratory motion estimation tasks and can handle the sliding motion correctly.

Each training image I_j is registered to the reference image I_R using the *deeds* algorithm, resulting in a dense motion field $\mathbf{u}_j$, which encodes the respiratory-related motion between I_j and I_R. Estimated dense motion fields are vectorized and concatenated into a data matrix $\mathbf{U} \in \mathbb{R}^{nV\times T}$, where nV is the number of image dimensions multiplied by the number of image pixels/voxels and T is the number of training images. Displacement vectors are the columns of the data matrix $\mathbf{U}$ and now a linear statistics can be performed on the data matrix to generate the motion model.

PCA is performed on $\mathbf{U}$ to find the orthogonal basis vectors (eigenvectors or principal components), which can describe the variation of motions that are uncorrelated to each other. To find the basis vectors, first the covariance matrix is generated from $\mathbf{U}$ and then an eigenvalue decomposition is performed on the covariance matrix $\mathbf{C}$:

$$\mathbf{C} = \hat{\mathbf{U}}^T\hat{\mathbf{U}} = \frac{1}{T}\sum_{t=1}^{T}(\mathbf{u}_j - \mu)(\mathbf{u}_j - \mu)^T = \mathbf{P}\boldsymbol{\Sigma}^{\mathbf{2}}\mathbf{P}^T \tag{3.3}$$

where $\hat{\mathbf{U}}$ is the mean-free data matrix, $\mu = \frac{1}{T}\sum_{j=1}^{T}\mathbf{u}_j$ is the mean displacement field, $\mathbf{P} \in \mathbb{R}^{nV\times nV}$ is the matrix containing eigenvectors of $\mathbf{C}$ in its orthonormal columns and $\boldsymbol{\Sigma}^2 \in \mathbb{R}^{nV\times nV}$ is a diagonal matrix containing the corresponding eigenvalues for the

eigenvectors. The eigenvectors span a linear space, and the eigenvalues in Σ represent the variance covered by the corresponding eigenvectors. The matrix $\mathbf{P}$ includes all eigenvectors, most of which containing irrelevant information or noise. Typically, the eigenvectors associated with the k largest eigenvalues are selected to obtain a compact model containing most relevant information. The number k can be chosen based on the variance threshold selected by the user and in this work, the variance threshold is chosen to be 95% [Preiswerk et al., 2014; McClelland et al., 2013; Klinder and Lorenz, 2012; Boye et al., 2013]. The reduced matrix $\mathbf{P_k} \in \mathbb{R}^{nV \times k}$ has orthonormal unit vectors in the columns that parameterizes the model space, the subspace of plausible mean-centered displacement fields. Using the basis vectors, any displacement fields $\mathbf{u}$ that belongs to this model space can be expressed as:

$$\mathbf{u} = \mu + \mathbf{P}_k \mathbf{\Sigma_k} \mathbf{c} \tag{3.4}$$

where $\mathbf{c} \in \mathbb{R}^{k \times 1}$ is a weight vector, containing weights for each basis. Using the motion model $\mathbf{P}_k\mathbf{\Sigma}_k$, a dense motion field $\mathbf{u}_t$ for time frame t can be reconstructed from the sparse motion field $\tilde{\mathbf{u}}_t$. An optimal weight vector $\mathbf{c}_t$ that best projects $\tilde{\mathbf{u}}_t$ into the model space can be found by minimizing the following regularized least-squares problem:

$$E(\mathbf{c}_t) = ||\tilde{\mathbf{P}}_k \mathbf{\Sigma}_k \mathbf{c}_t - (\tilde{\mathbf{u}}_t - \tilde{\mu})||_2^2 + \eta ||\mathbf{c}_t||_2^2 \tag{3.5}$$

where $\tilde{\mathbf{P}}_k\mathbf{\Sigma}_k$ and $\tilde{\mu}$ are the sparse versions of the model and the empirical mean and $\eta \geq 0$ is the weight parameter for the regularization term $||\mathbf{c}_t||_2^2$. The regularization term ensures plausibility of the reconstructed displacement field and prefers smaller deviation from μ by penalizing the weight vector $\mathbf{c}_t$ from having larger norms. With the estimated $\mathbf{c}_t$, the dense motion field $\mathbf{u}_t$ can be reconstructed using the Equation 3.4. This is a common way to reduce the effects of noise and sparsity on dense motion reconstruction, and was also used in [Klinder and Lorenz, 2012] and [Preiswerk et al., 2014].

3.2.3 Iterative coupled convex optimization

The cost terms in Equation 3.1 cannot be minimized simultaneously. The dissimilarity cost term is first minimized using the block-matching algorithm and based on the $\tilde{\mathbf{u}}_t$ obtained, the regularized dense motion field $\mathbf{u}_t$ can be reconstructed. In this way, however, the cost function cannot be minimized sufficiently, resulting in a suboptimal dense motion field estimation, which has regions that are not smooth. To aim a better estimation of the dense motion field, a joint optimization scheme that optimizes each cost term in alternating manner is adopted. For this an auxiliary vector $\tilde{\mathbf{v}}_t$ is introduced into the cost function modifying the Equation 3.1 into:

$$E(\tilde{\mathbf{u}}_t, \tilde{\mathbf{v}}_t) = \sum_{\Omega N} \mathcal{D}(I_R, I_M, \tilde{\mathbf{u}}_t) + \theta ||\tilde{\mathbf{u}}_t - \tilde{\mathbf{v}}_t||_2^2 + \alpha \mathcal{R}(\tilde{\mathbf{v}}_t). \tag{3.6}$$

In the modified equation, the dissimilarity term $\mathcal{D}$ and the regularization term $\mathcal{R}$ is decoupled and $\mathcal{R}$ does not depend on the sparse motion field $\tilde{\mathbf{u}}_t$ anymore. The extra coupling term is added, and it penalizes the deviation between the additional auxiliary vector $\tilde{\mathbf{v}}_t$ and $\tilde{\mathbf{u}}_t$. For $\theta \to \infty$, above equation is equivalent to the optimization of Equation 3.1. The control parameter α for the regularization term is implicitly selected by determining k, the number of eigenvectors of the model.

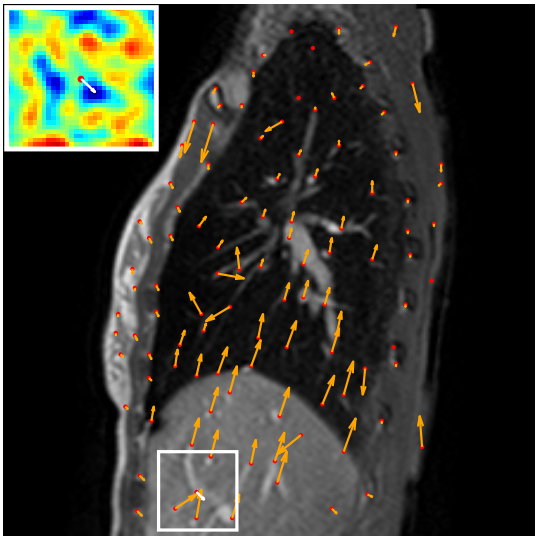
The initial optimization is performed using the block-matching algorithm without constraint by the coupling term, i.e. $\theta = 0$ and a sparse motion vector $\tilde{\mathbf{u}}_t$ can be determined. After obtaining $\tilde{\mathbf{u}}_t$, the regularized sparse motion field $\tilde{\mathbf{v}}_t$ can be determined by first, finding the appropriate weight vector $\mathbf{c}_t$ using the Equation 3.5 and second, solving the Equation 3.4 with the sparse model $\tilde{\mathbf{P}}_k\mathbf{\Sigma}_k$ and the estimated $\mathbf{c}_t$. As mentioned earlier, the cost map for each keypoint is stored, and now it can be updated after the determination of $\tilde{\mathbf{v}}_t$ by adding the coupling term $\theta||\tilde{\mathbf{u}}_t - \tilde{\mathbf{v}}_t||_2^2$ to the stored values. With the updated cost map, a new sparse motion field $\tilde{\mathbf{u}}_t$ can be computed by selecting the point with the minimal cost, then again $\tilde{\mathbf{v}}_t$ is determined using the sparse model and it is performed alternately until an optimal sparse displacement field can be found and the entire cost converges into the minimum value. We use gradually increasing values for parameter θ and the cost function converges quickly after several iterations.

As shown in Figure 3.3a, the displacement field estimated via block-matching under unconstrained condition contain many inconsistent displacement vectors, which cannot describe the physiological motion properly. The cost map of an example keypoint shown upper left corner contains several local minima, which lead to a suboptimal result. The cost map is smoothed after the first update made after the regularization using the sparse motion field as shown in Figure 3.3a. Most of the displacement vectors are consistent in direction after the regularization and the cost map is much smoother than the initial cost map, however, as marked with red circles, there are still inconsistent vectors. The result after the sixth iteration (Figure 3.3c), all displacement vectors are consistent and comparable to the ground truth displacement vectors computed using the *deeds* algorithm (Figure 3.3d).

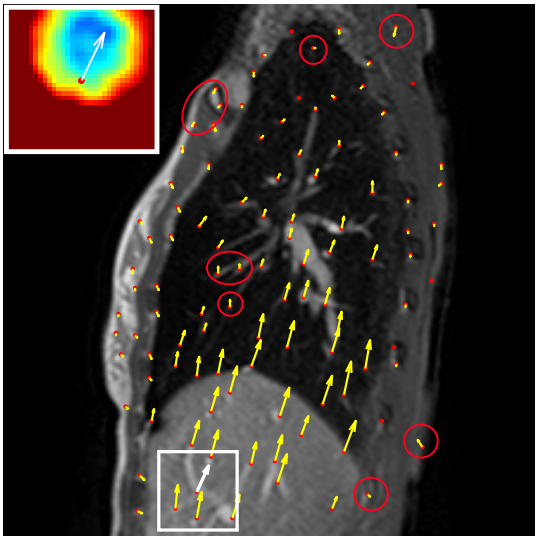
To ensure temporal consistency, an additional coupling term can be used that incorporates the prior knowledge on previous motion. The Equation 3.6 can be modified as:

$$E = \sum_{\Omega N} \mathcal{D} + \theta||\tilde{\mathbf{u}}_t - \tilde{\mathbf{v}}_t||_2^2 + \beta||\tilde{\mathbf{v}}_t - \tilde{\mathbf{u}}_{t-1}||_2^2 + \alpha\mathcal{R} \tag{3.7}$$

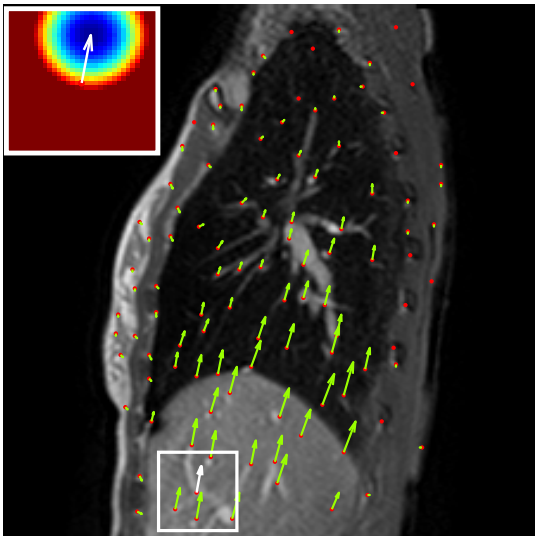
where the additional coupling term penalizes the deviation of estimated displacement field from the displacement of previous time point $t-1$. With the Equation 3.7, the resulting motion estimation is also temporally smooth.



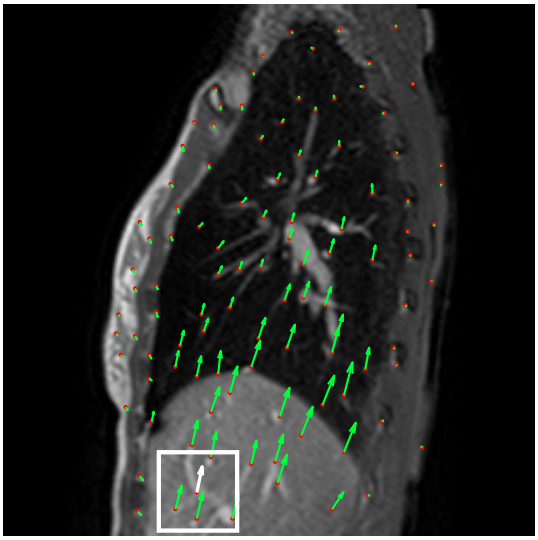
(a) Displacement vectors after the block-matching



(b) Displacement vectors after the first regularization



(c) Displacement vectors after the sixth iteration



(d) Ground truth displacement vectors from *deeds* algorithm

Figure 3.3: Example image of spare motion field estimation results. The displacement vectors are shown with arrows and in the top left corner the cost map for the keypoint marked with white box and white arrow is displayed. The displacement vectors estimation after the initial block-matching (a) have some inconsistent estimations, as marked with white arrow. After the regularization (b), some vectors are corrected, however, there are still some inconsistencies as marked with red circles. The displacement estimation after six iterations (c), the result looks similar to the ground truth estimation (d) from *deeds* algorithm. (Image source: [Ha et al., 2018])

3.3 Experiments and Results

The experiments are performed on three publicly available 2D and 3D datasets of MRI and US images. In the following, some details on each dataset are provided, as well as little information on experimental setting.

2D+t MRI Four sequence 2D MRI images of thorax from [Baumgartner et al., 2017], each from a different patient, is used for evaluation. The images are T1-weighted and each dataset (VolA-VolD) consists of 40 sagittal frames taken under free-breathing and one sagittal frame taken under breath hold. Images have a pixel resolution of 215×288 with an isotropic pixel spacing of 1.39 mm. From each dataset, we select two sagittal slices for evaluation, one at the left lung center and the other at the right lung center. Corresponding landmarks (10-20) are manually selected in the lung and liver region of five random 2D images at different time point. For this dataset, we use the first half of the sequence as training data to generate a patient-specific motion model and the rest as the test images.

4D MRI Two 4D MRI datasets (*sl010* and *sl014*) of the thorax/abdomen from [Boye et al., 2013] are used for evaluation. Each dataset is from a different patient and contains 200 frames of 3D MRI images taken under free-breathing condition, which includes several respiratory cycles. The images have a voxel resolution of $224 \times 224 \times 50$ and $224 \times 224 \times 52$, an isotropic voxel spacing of 1.21 mm and 1.30 mm for the sagittal plane and an inter-slice distance of 5 mm respectively. Due to a large inter-slice distance, only sagittal in-plain motion is estimated. For evaluation, 27 corresponding landmarks were set manually in the lungs and liver in 10-11 randomly selected image frames. We divided the dataset into three folds, using one fold for training data and the remaining two folds as test data. For keypoints extraction, three slices including left and right lungs and heart are selected (slice number of 16, 22, 34).

4D US Nine 4D US dataset (*SMT01-09*) of liver, each containing 92-96 3D image frames from the CLUST Challenge [De Luca et al., 2015] is used for 3D motion estimation. The images have a voxel resolution of $227 \times 227 \times 229$ voxels with an isotropic voxel spacing of 0.7 mm, and the temporal resolution is 8 Hz. We set 1-2 landmarks manually in 9-10 randomly selected frames of each dataset for evaluation. Since the frequency of the image acquisition of this dataset is higher than the MRI datasets, we divide the dataset by the frame number, using the odd frames as training data and the even frames as test data.

In all experiment, the first frame is selected as the reference image, to which the other image frames are registered. The generation of motion model is also performed by registering the rest image frames to the reference frame using *deeds* algorithm [Heinrich et al., 2013a].

The experiments are performed to first determine optimal parameters and evaluate the effect of each component of our framework. For these experiments, we use a subset of 4D MRI dataset and once the optimal parameters are determined, an exhaustive evaluation is performed using these parameters with 2D+t, 4D MRI datasets and 4D US dataset. Moreover, we compare the results of 2D+t and 4D MRI datasets with a state-of-the-art method for 2D/3D image registration named *RealTITracker* from [Zachiu et al., 2015]. We use mean TREs as an evaluation method, which is calculated based on the manually set landmarks and perform paired t-tests with a significance level of 5% ($p < 0.05$) by paring the patient-specific mean TREs to assess the statistical significance of the mean TRE difference between compared methods.

3.3.1 Effect of different parameters

For evaluation of different components of the proposed methods, a subset of 4D MRI dataset *sl010* is used. A patient-specific motion model is generated using the first third of frames (*training data*) of the image sequences and the rest of the images are treated as the intraoperative online images (*test data*), which are normally taken during the treatment. In this work, the number of image frames acquired during the treatment is determined based on the first clinically available MRI-guided radiation therapy system (Viewray), which was reported to be able to acquire three parallel frames simultaneously at 2 fps rate [Mutic, 2012].

Feature points and block-matching In the proposed method, estimation of displacement vectors using a block-matching algorithm and the sparse motion model is performed on a set of sparse landmarks extracted by the Harris corner detector. This number can be controlled by the radius size of the non-maximum suppression algorithm and has an influence on the registration accuracy and the computation time of the framework. To assess its effect, we experiment with the different radii of non-maximum suppression algorithm, i.e. with a different number of keypoints. For this experiment, we set fix the other hyperparameters such as the block-size, size of the searching window and the control parameter of the coupling term θ. The block-size of block-matching algorithm is set to 11 pixels and the size of the search region is determined according to the maximal distance observed in automatic registration of the training data. We set

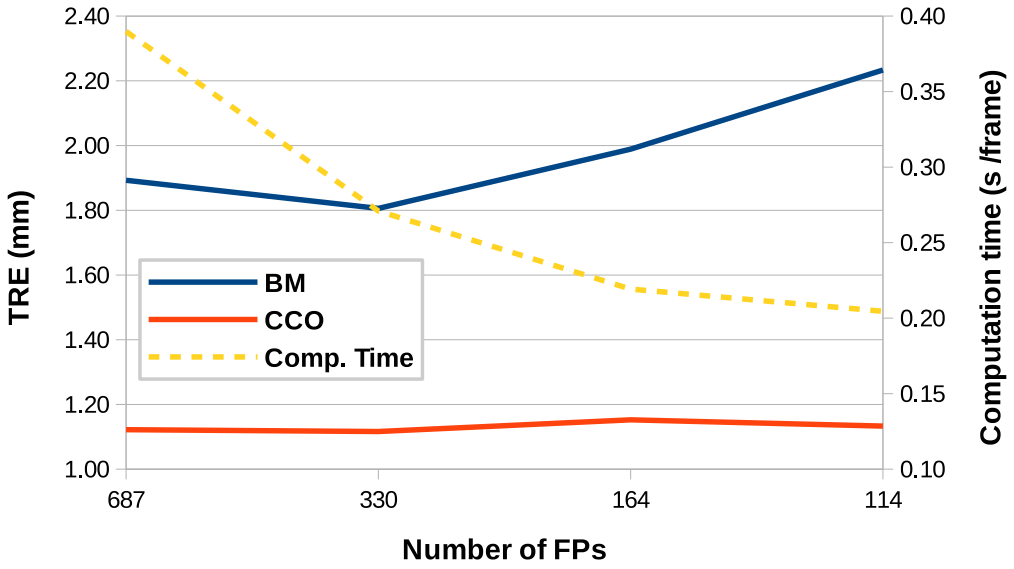


Figure 3.4: Mean TREs of the reconstructed dense motion fields from the estimated sparse motion field with the different number of keypoints. The keypoints are extracted from the online images (three slices) and the final accuracy is calculated after the block-matching and 6 iterations of coupled convex optimization for 27 landmarks from each of 10 selected frames. Computation time includes the duration of block-matching and six iterations of coupled convex optimization. The influence of the sparsity (the number of keypoints) is small when the coupled convex optimization is employed and the sparse-to-dense extrapolation can be done with small number of keypoints. (Image source: [Ha et al., 2018])

the control parameter of the coupling term θ to logarithmically increasing values from 0.03 to 1.

As shown in Figure 3.4, the mean TRE of the estimated dense motion field tends to decrease as the number of the keypoints increases (blue line), if only the block-matching is performed. However, with the coupled convex optimization, which incorporates the patient-specific motion model, the effect of the sparsity is not minimal (orange line). The computation time decreases as the number of keypoints decreases and given the application of the coupled convex optimization algorithm, approximately 100 keypoints can be used for the estimation of the sparse displacement fields.

Coupled convex optimization Comparison between the different number of iterations for the coupled convex optimization is shown in Figure 3.5. The accuracy of the

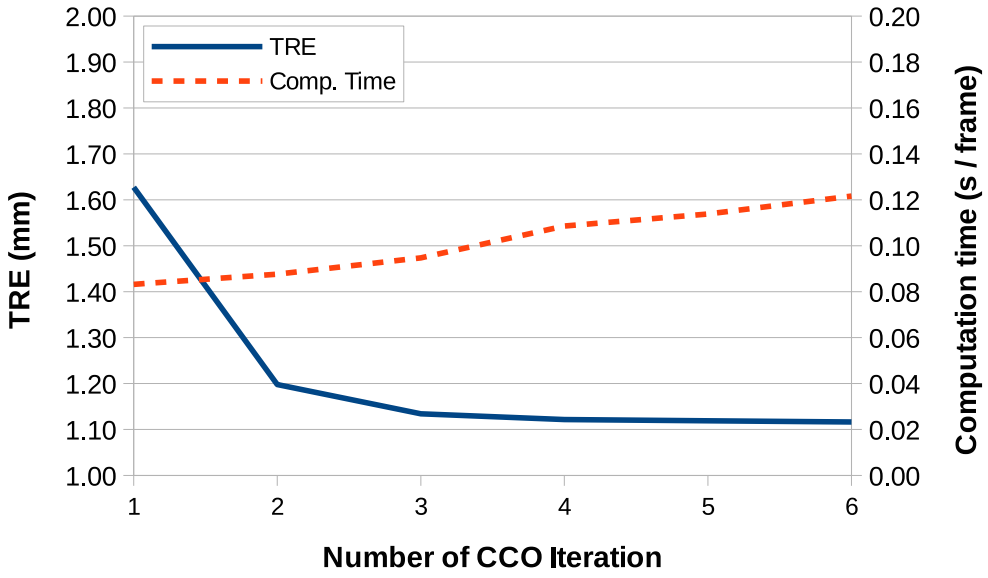


Figure 3.5: Mean TREs (blue line) of reconstructed dense displacement field estimated using a different number of iterations for coupled convex optimization. TRE is computed for 27 landmarks set on 10 image frames, and 330 key-points are used for the block-matching and for sparse motion model. The increase in the computation time of coupled convex optimization (dotted orange line) is not significant with the increasing number of iterations. (Image source: [Ha et al., 2018])

result is represented with the mean TREs shown in the blue line, and the computation time for the different number of iterations is also given in the dotted orange line.

With one iteration, it is the same as performing the model-based regularization on the block-matching result and comparable to the work of Preiswerk et al. [2014]. The accuracy increases with the increasing number of iterations, while the increase in the computation time is minimal (only 40 ms for increase in the number of iterations from one to six).

3.3.2 Evaluation on MRI data

A quantitative evaluation of our method is performed using two 4D MRI datasets (*sl010* and *sl014*) and four 2D+t MRI datasets (*volA*-*volD*). We compare the result of the proposed method with the gold standard algorithm (*deeds*), which was used to generate the patient-specific motion model. Moreover, we compare our result with the *RealTI-Tracker (Realtime Image-based Tracker)*, a state-of-the-art method for 2D/3D optical

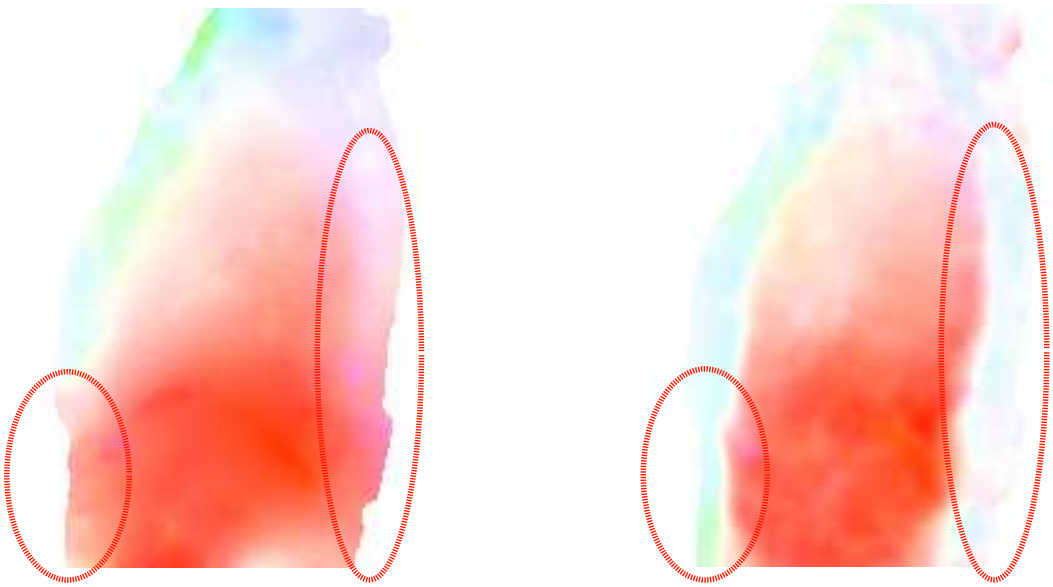
Table 3.1: Experimental results of 2D+t MRI datasets: TRE (mean landmark distance in mm) and computation time (ms/frame) of *deeds* registration, block-matching (BM), regularization after block-matching ($\eta > 0$) and our proposed method with and without temporal component are given.

	volunteerA	*volunteerB*	*volunteerC*	*volunteerD*	mean (mm)	CPU (ms/f)	GPU (ms/f)
before	6.26 ± 3.47	3.47 ± 1.85	8.09 ± 2.55	5.65 ± 3.65	5.87	–	–
deeds	0.97 ± 1.00	0.59 ± 0.31	0.73 ± 0.40	0.71 ± 0.69	**0.75**	$\approx \mathbf{9.2 \cdot 10^3}$	–
BM ($\eta = 0$)	1.77 ± 0.99	0.97 ± 0.33	1.32 ± 0.44	1.75 ± 0.68	1.45	≈ 100	≈ 1
BM ($\eta > 0$)	1.60 ± 1.41	0.92 ± 0.44	1.32 ± 0.68	1.65 ± 1.23	1.37	≈ 101	≈ 6
proposed	1.13 ± 1.27	0.73 ± 0.46	0.87 ± 0.72	1.01 ± 1.11	**0.94**	$\approx$ **114**	$\approx$ **3**
proposed (w/ temp)	1.27 ± 0.99	0.72 ± 0.42	0.85 ± 0.47	0.96 ± 0.76	0.95	≈ 118	≈ 4
RealTITracker	1.12 ± 0.88	0.82 ± 0.37	0.83 ± 0.74	1.37 ± 1.19	**1.04**	$\approx$ **310**	–

flow based medical image registration algorithm, which is developed for MRI-guided realtime tracking of tumor/organs [Zachiu et al., 2015]. We use the *RealTITracker* with its *PCAMotionDescriptor* add-on, which implements a PCA-based motion descriptor to ensure the spatio-temporal coherency of the complex organ deformation through a learning step [de Senneville et al., 2015]. The registration using *RealTITracker* is performed as 2D multislice registration with *PCAMotionDescriptor* add-on, where we use four PCA basis as recommended in the paper [de Senneville et al., 2015].

For the evaluation, the hyperparameters determined in the previous experiment (Section 3.3.1) are used. For both 4D and 2D+t dataset, the block-size was set to 11 pixels and the search radius was set to 16 and 15 pixels respectively. Noise estimate σ is set to 0.6 and the neighborhood distance $\delta = 2$. The number of iterations for coupled convex optimization was chosen to be six and for the control parameter θ, logarithmically increasing values between 0.03 to 1 were used. For 2D+t MRI dataset, an additional temporal constraint was employed using the second regularization term to obtain temporally smooth dense motion fields (see Equation 3.7). The parameter β was set to 0.5 and 0.25 respectively for the first and the second iterations and set to 0 from the third iteration.

The quantitative results are presented in Table 3.1 and Table 3.2. The mean TRE values as well as its standard deviation of the original landmark distances (without registration), the gold standard registration method (*deeds* registration), the reconstructed dense motion field directly after the block-matching (BM ($\eta = 0$)) and after a spatial



(a) motion field (*RealTITracker*) **(b)** motion field (proposed)

Figure 3.6: Example dense deformation field estimation of same slice using *RealTITracker* (a) and proposed method (b).The cranio-caudal motion in the marked regions (red circles) including skin and bones should be close to zero, which is correctly estimated with the proposed method. (Image source: [Ha et al., 2018])

regularization with a motion model (BM ($\eta > 0$)) and our proposed method with the optimization using Equation 3.5 and Equation 3.7 are given for each dataset. We have also performed paired t-test to determine the significance of the difference between the results of different methods. Computation times for each method is also given in seconds per frame for CPU computation and GPU computation.

The t-test result indicate that the difference between the mean TREs of the gold standard algorithm (*deeds*) and the proposed method is insignificant, however, the differences of mean TREs are significant between the proposed method and the block-matching with and without the spatial regularization (BM ($\eta = 0$) and BM ($\eta > 0$)).

The comparison to the *RealTITracker* with *PCAMotionDescriptor* also shows that our method has a significantly better accuracy. Specifically, our approach has a clear advantage in estimating sliding motion. In Figure 3.6, a qualitative result is shown to compare the proposed method and *RealTITracker*. While our method correctly estimate sliding motion at the borders between rib cage and lung and belly skin and liver, the compared method cannot estimate sliding motions as shown in the Figure 3.6. We also have performed a quantitative analysis on the ability to handle the sliding motion of all tracking approaches presented in the Table 3.1. This is done by analyzing the

Table 3.2: Experimental results of 4D MRI datasets: TRE (mean landmark distance in mm) and computation time (ms/frame) of *deeds* registration, block-matching (BM), regularization after block-matching ($\eta > 0$) and our proposed method without temporal component are given.

	sl010	*sl014*	Mean (mm)	CPU (ms/f)	GPU (ms/f)
before registration	3.97 ± 2.33	1.78 ± 1.04	2.87	–	–
deeds registration	1.12 ± 0.62	1.01 ± 0.52	**1.06**	$\approx \mathbf{6 \cdot 10^4}$	–
BM ($\eta = 0$)	1.81 ± 1.16	1.54 ± 0.94	1.72	≈ 100	≈ 27
BM ($\eta > 0$)	1.63 ± 1.02	1.26 ± 0.66	1.45	≈ 127	≈ 36
proposed method	1.12 ± 0.64	1.07 ± 0.57	**1.10**	$\approx$ **204**	$\approx$ **34**
RealTITracker	2.07 ± 1.29	1.67 ± 0.92	**1.87**	$\approx \mathbf{3.8 \cdot 10^3}$	–

landmarks additionally set on the bony structures (ribs and vertebrae). The landmarks on the vertebrae had mostly no motion, and the landmarks set on the ribs had the motion directed orthogonally to the predominant cranio-caudal motion of the lungs and livers.

The registration result of 2D+t datasets presented in Table 3.2 show that the proposed method is significantly better than just using block-matching (BM ($\eta = 0$)) or a motion model based regularization after the block-matching (BM ($\eta > 0$)). Moreover, the motion compensation rate[1] of the proposed method ($\approx 84\%$) is only marginally different to that of the *deeds* registration ($\approx 87\%$).

In addition to the accuracy of the motion estimation, the computation time is an important aspect of the proposed method. The computation times of each method is shown in both Table 3.1 and Table 3.2 for CPU and GPU computation, which shows the realtime ability of the proposed method. Compared to the *deeds* registration algorithm, which takes about a minute to register an image frame, our approach is approximately 300 times faster even on the CPU, taking only approximately 200 ms for the registration.

3.3.3 Evaluation on ultrasound data

Registration of ultrasound images are usually more challenging than registration of images such as MRI or CT, because of its lower image quality compared to the other modalities. Moreover, due to its smaller field-of-view and the use of the probe, which might not be fixed to one position during the image acquisition, some anatomical struc-

[1]The motion compensation rate is defined as $(A-B)/A*100$, where A is the mean Euclidean distance of landmarks $\mathbf{x}, \mathbf{y} \in I_R, I_M$ and B is the mean Euclidean distance of $\mathbf{x}, \mathbf{y}' \in I_R, (I_M + \mathbf{u})$.

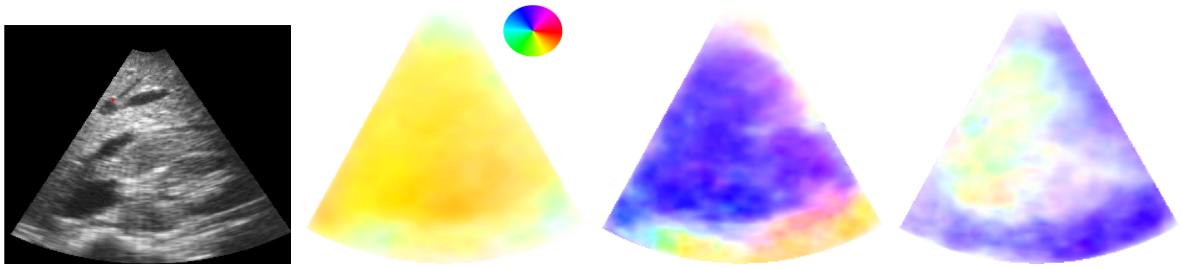


Figure 3.7: US reference image with a ground truth landmark marked with a red point and the first three principal components of the model.

tures might not be clearly visible or might disappear from the image field-of-view, making it more difficult to track structures of interest.

In this experiment, we use 4D US datasets (*SMT01-09*) from CLUST Challenge [De Luca et al., 2015]. The evaluation was done using the landmarks we set manually on each dataset. Example images of the reference image with a ground truth landmark as well as the first three principal components of the model is depicted in Figure 3.7. The hyperparameters for this experiment were determined empirically, where we use the block-size of 7 pixels and different searching windows for each dataset. The parameter θ for coupled convex optimization was also set as logarithmically increasing values between $0.02 - 5.0$ and β for the temporal regularization is set as in the previous experiment (see Section 3.3.2).

In Table 3.3, TREs for different configurations are summarized. The proposed methods with and without temporal regularization significantly outperform the block-matching results (BM ($\eta = 0$)) and the results of model-based regularization after the block-matching (BM ($\eta > 0$)) in accuracy. The adaptation of temporal regularization improves the accuracy, however, the improvement is not significant. Although the accuracy of *deeds* registration method was significantly better than the proposed method, the motion compensation rate was comparable (*deeds*: $\approx 78\%$, ours: $\approx 77\%$) and the mean TRE of 2.20 mm is also similar to the average inter-observer distance of 1.67 mm.

The proposed method can be computed significantly faster than *deeds* algorithm, however, for 3D dataset it still can not be computed in realtime. Since the imaging frequency of the used dataset is 8 Hz, the computation time should be under 125 ms to achieve realtime performance. The most time-consuming part of the proposed method is the block-matching and coupled convex optimization, which can be improved by the implementation on GPU using convolution as introduced in the following.

Table 3.3: Experimental Result of 4D US datasets: TRE (mean landmark distance in mm) and computation time (s/frame) of *deeds* registration, block-matching (BM), regularization after block-matching ($\eta > 0$) and our proposed method with and without temporal component. (from [Ha et al., 2018])

Data set (landmark)	without registration	*deeds* registration	BM ($\eta = 0$)	BM ($\eta > 0$)	Model-based w/o temporal	Model-based w/ temporal
SMT01 (1)	7.86 ± 3.38	0.93 ± 0.18	2.63 ± 1.25	1.60 ± 1.21	1.04 ± 0.33	0.99 ± 0.30
SMT01 (2)	8.31 ± 2.45	1.35 ± 0.73	2.56 ± 0.89	1.92 ± 0.65	1.44 ± 0.86	1.43 ± 0.85
SMT02 (1)	6.21 ± 2.63	0.88 ± 0.09	2.65 ± 1.67	1.05 ± 0.38	0.90 ± 0.19	0.90 ± 0.19
SMT03 (1)	6.07 ± 4.33	0.72 ± 0.22	3.76 ± 1.87	1.95 ± 2.11	1.20 ± 1.10	1.04 ± 0.68
SMT04 (1)	14.20 ± 7.44	6.18 ± 5.51	9.91 ± 5.26	6.76 ± 5.44	6.52 ± 5.46	6.63 ± 5.35
SMT05 (1)	12.30 ± 7.01	3.05 ± 1.42	6.23 ± 1.94	5.46 ± 3.32	4.30 ± 2.87	3.24 ± 1.40
SMT06 (1)	15.91 ± 5.66	2.07 ± 0.73	6.50 ± 4.21	4.40 ± 3.29	3.14 ± 2.40	2.35 ± 1.00
SMT06 (2)	14.16 ± 7.99	4.64 ± 3.04	8.84 ± 3.69	7.59 ± 6.26	6.44 ± 5.72	4.73 ± 3.72
SMT07 (1)	5.35 ± 2.20	1.12 ± 0.52	1.85 ± 0.71	1.35 ± 0.58	1.19 ± 0.43	1.19 ± 0.47
SMT08 (1)	4.20 ± 2.63	0.63 ± 0.21	1.84 ± 0.78	0.70 ± 0.32	0.63 ± 0.19	0.62 ± 0.20
SMT09 (1)	9.46 ± 3.01	1.08 ± 0.65	3.10 ± 1.65	1.51 ± 0.99	1.19 ± 0.76	1.14 ± 0.69
mean (mm)	9.46	**2.06**	4.53	3.12	**2.54**	**2.20**
CPU (ms/f)	–	$\approx \mathbf{6 \cdot 10^4}$	≈ 160	≈ 265	≈ 612	$\approx$ **610**
GPU (ms/f)	–	–	≈ 20	≈ 42	≈ 109	$\approx$ **116**

3.3.4 GPU implementation

When using a standard block-matching algorithm computed on CPU (Intel Core i5-6600@ 3.3 GHz (4 cores) with 32 GB RAM), the full reconstruction of the dense motion field takes up to 5 seconds using the proposed method. By implementing the algorithm on GPU, the computation time can be reduced to meet realtime requirement. The SSC descriptor computation, block-matching algorithm, coupled convex optimization as well as the motion field reconstruction can be performed on the GPU reducing the computation time up to approximately 6 times for 4D datasets and 30 times for 2D+t dataset. In addition to the GPU computation, we optimized the computation time for the block-matching algorithm by utilizing convolutional filters to enable parallelization of the local image dissimilarity computation.

The dissimilarity cost of the proposed method is calculated using the sum of squared distances (SSD) between two SSC descriptor patches $P \in \mathbb{R}^{\Omega_p}$ and $Q \in \mathbb{R}^{\Omega_s}$ as follows:

$$\mathcal{D}(d_x, d_y) = \frac{1}{\Omega_p} \sum_{i,j \in \Omega_p} P_{i,j}^2 - 2P_{i,j}Q_{i+d_x,j+d_y} + Q_{i+d_x,j+d_y}^2 \tag{3.8}$$

where Ω_p denotes the subset of pixels in the image patch, Ω_s the subset of pixels in the corresponding patch from the search region and $P_{i,j}$ and $Q_{i+d_x,j+d_y}$ are the SSC descriptor at pixel coordinate (i, j) and $(i + d_x, j + d_y)$ respectively. (d_x, d_y) is the displacement vector defined within the search region. The first and last term of the Equation 3.8 can be computed easily by performing element-wise multiplication of the matrices, whereas the computation of the second term cannot be computed in the same way. To accelerate the computation time, we utilize convolution to compute the second term, where the reference patch P is set as the convolution filter and applied on the search region patch Q. With an appropriate reshaping of the matrices, the computation can be parallelized for the entire image, and we can compute all three terms using convolutions for the whole image at once. On the CPU, this new approach reduces the computation time to under 1 second and on the GPU it can be computed in realtime, reducing the computation time more than one order of magnitude as shown in Table 3.2 - Table 3.3.

3.4 Discussion

With our experiment, we have shown that given a highly accurate registration algorithm such as *deeds* [Heinrich et al., 2013a], which enables the generation of a high quality motion model, the proposed method is able to accurately estimate tumor and/or organs at risk and furthermore, handle the sliding motion.

The quantitative evaluation is performed on three publicly available datasets that represent the use-case of the image-guided interventions with MRI and ultrasound images. The result shows that the proposed method has nearly reached the lower-bound of the gold standard dense registration algorithm in accuracy 100x faster by utilizing a small number of keypoints and joint optimization of the cost function. The speed-up was enabled by the use of GPU implementation and the parallelization of dissimilarity computation using convolutions. In the comparison with the state-of-the-art optical flow registration method [Zachiu et al., 2015], our method was advantageous in accuracy and computation time, setting the new state-of-the-art in realtime motion estimation.

Compared to our previous work [Wilms et al., 2016], more detailed experiments on each module is performed in this work to analyze the impact of the number of keypoints and the number of iterations for the coupled convex optimization. It is shown that the sparsity of the model represented by the number of keypoints can be significantly increased, thereby reducing the number of required keypoints, when the alternating optimization scheme via coupled convex optimization is used. Our experiments show that the previous work [Preiswerk et al., 2014], which separately performs the regularization on the block-matching result, is not sufficient enough and can be improved by the proposed method significantly.

The computation time of the proposed framework is also significantly reduced by GPU implementation of the descriptor computation, block-matching algorithm, coupled convex optimization and motion field reconstruction. Especially by optimizing block-matching algorithm using convolutional filters and parallel computing on GPU, the computation time was reduced more than one order of magnitude. The same algorithm is also used in Chapter 5 to determine initial displacement vectors based on learned features.

Especially for image-guided interventions, the proposed approach has a great impact and can provide a viable solution for the systems with limited online image acquisition time such as MRI-Linac by enabling a realtime motion estimation, which can be realized by sparse-to-dense motion field reconstruction. With the improved accuracy in target tracking, the safety margins such as internal target volume (ITV) or planning target volume (PTV) can be reduced in applications such as image-guided radiation therapy. As a result, the dose delivery can be improved, sparing more healthy tissues around the target and possibly speed up the therapy. It is stated in the recommendations for the implementation of a realtime tracking response to respiratory motion of AAPM Task Group 76 report [Keall et al., 2006], that the total time delay of a realtime tracking systems should be kept as short as possible and, in any case, not more than 0.5 seconds. The experiments on all three datasets resulted in the computation time of under 150 ms ($<$50 ms for MRI-based tracking and $<$120 ms for US-based tracking), which is consider-

ably below the required threshold. This leaves enough time for the remaining processes in the pipeline such as image acquisition and beam re-positioning to be performed.

In current implementation, the variability of respiratory motion of patient in planning phase and treatment phase is not taken into account. If the difference of respiratory motion pattern in pre-treatment and treatment phase is large, current implementation might not be able to estimate the motion accurately, since search region of the block-matching algorithm is selected depending on the respiratory motion in pre-treatment phase. Possible improvements that can be made in the future work are the online adaptation of the motion model and block-matching algorithm to deal with the problems such as baseline shifts [Ruan et al., 2009] and integration of temporal information in patient-specific model using e.g. spatio-temporal PCA to replace the temporal constraint presented in this work.

3.5 Summary

In this chapter, a novel approach to a patient-specific model based motion compensation framework for image-guided interventions without invasive markers are presented. The proposed method incorporates the information on the respiratory motion of the patient obtained in the pre-treatment phase to enable motion estimation for the whole image field of view from an incomplete image data available during the course of treatment. By jointly optimizing a GPU-accelerated block-matching of sparse keypoints and a patient-specific motion model for regularization, the proposed method achieves comparable accuracy of a state-of-the-art registration approach in few milliseconds, which is 100 times faster than the compared gold standard method.

The algorithmic contribution of the individual component of our framework is shown through detailed experiments. We have also validated our method using extensive experiments on three different MRI and US datasets of lungs and liver, which are largely influenced by the respiratory motion and therefore require accurate motion compensation. The results show that the proposed method significantly improves the accuracy compared to the approach that is based on a disjoint optimization of the cost term using single regularization step. Moreover, our method achieves better accuracy and faster computation time compared to the state-of-the-art optical flow based registration method [Zachiu et al., 2015].

The major drawback of the proposed approach is the requirement of ground truth deformation fields. Unlike the other annotations such as segmentation or landmarks, deformation fields cannot be generated manually by experts and requires a proper algorithm that can register images automatically. Thus, to acquire ground truth deformation fields, a gold standard image registration algorithm with suitable accuracy and robustness is

essential. There are some approaches that utilize synthetic ground truth deformations, however, these are mostly limited to simple deformations and cannot guarantee to appropriately represent the variations of deformation present in the data. To avoid the problem of acquiring ground truth deformation fields, one can think of utilizing simpler annotations such as segmentation or landmarks to incorporate auxiliary information into the registration approach, as presented in the next chapter.

Chapter 4

Weakly-Supervised Image Registration using Segmentation

In this chapter, we present an end-to-end learning based registration framework developed in the article Ha et al. [2020] published in the *Sensors*, an international peer-reviewed open access journal. In addition to the experiment performed in Ha et al. [2020], we perform an additional experiment in this chapter. The proposed framework is trained using a weak supervision based on auxiliary data, i.e. segmentation/labels, that provide semantic information for relevant structures in the images to be registered. Using a trained model, the framework based on machine learning can estimate a deformation field for a pair of images within half a second and without any manual interaction. We demonstrate improved accuracy and robustness of estimated deformation fields by incorporating auxiliary data. We first evaluate our framework on a real world 2D image dataset, for which we have a large number of labelled training samples. The framework is then evaluated on a 3D medical dataset with a smaller number of training samples, and we achieve a comparably good result to the state-of-the-art approaches.

4.1 Introduction and Related Works

While deep learning has shown its advantages over classical approaches in tasks such as medical image classification and segmentation, it still has not yet reached the accuracy of the state-of-the-art classical approaches in most 3D image registration tasks. Challenges for medical image registration using deep learning lie particularly in the limited number of training data, complexity of the task and the difficulties in acquiring ground truth data for supervised training. Especially the ground truth data, i.e. deformation fields between two images, is more difficult to generate compared to labels, segmentation masks or landmarks needed for classification or segmentation tasks.

One way to obtain ground truth data is to generate deformation fields using an existing registration method. In *SVF-Net* [Rohé et al., 2017], stationary velocity fields (SVF) are generated and used as the ground truth deformation fields to train a model to perform one-to-one registration between MR cardiac sequences.

When using an existing registration method, the quality of the generated ground truth deformation fields are strongly influenced by the accuracy of the used method and consequently affecting the accuracy of the final result. This was also one of the limitations in our previous work presented in Chapter 3, where a motion model was computed based on the deformation fields generated using an existing registration algorithm.

Instead of generating deformation fields for existing data, some approaches generate synthetic dataset, which consists of artificially generated deformation fields and images deformed using those deformation fields. Uzunova et al. [2017] generate random deformations based on the locality-based shape and appearance model, which served as ground truth deformations for images and landmarks generated using them. Krebs et al. [2018] generate synthetic ground truth deformation fields based on the registered ROI with only a small number of real ground truth deformation fields, and Sokooti et al. [2017] generate synthetic ground truth deformation fields randomly within a real displacement range.

Although the use of data augmentation with synthetically generated deformation fields can alleviate the problem of arranging ground truth deformation fields, randomly generated deformation fields might not represent plausible deformation present in the real dataset, and it is a cumbersome to generate deformation fields for training data. Furthermore, varying contrast is additionally challenging to synthesize, which could lead to a too simplistic training dataset.

To circumvent the problem of generating deformation field, auxiliary information can be incorporated to enable an indirect (weak) supervision of the training. Annotations such as organ/tumor segmentation and/or landmarks of anatomical features are easier to obtain for medical data compared to deformation fields. Many recent research works on medical image registration incorporate segmentation or landmarks into the training process of their deep learning frameworks to improve accuracy and robustness of the method [Hu et al., 2018c; Qin et al., 2018; Balakrishnan et al., 2019]. Hu et al. [2018c] introduced *Label-Reg*, which utilizes labels to train a U-Net like registration network. In their approach a set of anatomical labels are used, which are warped using the spatial transformer layer and the difference between warped moving labels and fixed labels are minimized during training. The spatial transformer layer is a differentiable module that spatially transforms the input during a single forward pass and outputs a transformed output. A localization network of the spatial transformation layer predicts parameters for spatial transformation that should be applied to the input feature map and creates

a sampling grid based on these parameters. Then, a differentiable warping is performed by sampling the input at the created sampling grid point to spatially transform the input. In VoxelMorph [Balakrishnan et al., 2019], segmentation can be used as auxiliary information to improve the accuracy. Qin et al. [2018] proposed a two-branched framework consisting of both a multiscale recurrent motion estimation branch and a segmentation branch, which share weights similar to Siamese networks. The output of the segmentation branch is warped using the output deformation field of the motion estimation branch, and a categorical cross-entropy loss is computed between the warped segmentation and target. In the above-mentioned approaches, the results can be biased depending on the labels used during training, and the label bias should be considered carefully.

4.2 Proposed Method

In the previous chapter, we present a statistical model based classic image registration approach, where ground truth deformation fields are used to generate a patient-specific motion model. Given displacement vectors for a sparse set of keypoints computed based on dissimilarity of hand-crafted image features, the statistical motion model was able to infer an optimal parameter set to reconstruct the most probable dense deformation field.

With GPU programming and efficient block-matching algorithm (3.3.4), the presented approach has achieved high accuracy and fast computation time. However, the necessity of a highly accurate gold standard registration algorithm is a significant drawback that limits this approach, since the accuracy of the approach inevitably relies on the accuracy of the gold standard registration algorithm. The approach presented in this chapter uses deep learning to overcome this limitation, by training CNN networks that do not require ground truth deformation fields. The fast computation time is guaranteed for inference time by the use of trained models and GPU computation. We train a feature network as well as registration networks to estimate deformation field, while explicitly utilizing semantic information provided by segmentation in feature extraction of individual images.

To estimate a deformation field, which best aligns the fixed image $I_f : \Omega \rightarrow \mathbb{R}$ to the moving image $I_m : \Omega' \rightarrow \mathbb{R}$, a straightforward U-Net [Ronneberger et al., 2015] and one or two encoder-like registration networks are utilized to build a deep learning registration framework. Image domain Ω and Ω' are the regions of interest or the image field of view, which have the same pixel/voxel resolution ($m \times n$ for 2D case and $m \times n \times o$ for 3D case) in this case. The U-Net is trained to estimate segmentation labels of important organs, which should be aligned. For the registration part, two networks with the same

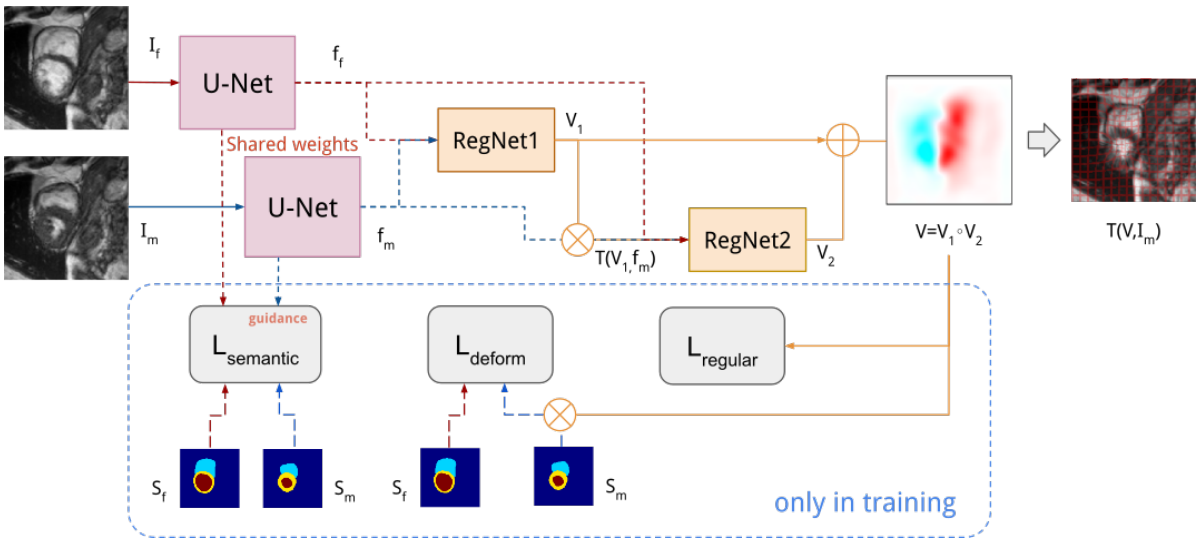


Figure 4.1: Overview of the semantically-guided registration framework: Dataflow of fixed image is represented with red lines and the moving image with blue lines. The yellow lines indicate the dataflow of deformation field. The dotted lines indicate segmentation labels. The circled multiplication symbol is used to represent the differentiable warping operation, whereas for the circled plus symbol is used to indicate addition of displacement fields. Segmentation labels are only required during training for computation of losses.

network architecture are used and trained sequentially. The output deformation field from each registration network has half the size of the input and is upsampled to match the input size before spatially transforming the segmentation (spatial transformation) [Jaderberg et al., 2015]. As shown in the graphical overview in Figure 4.1, losses are computed based on the segmentation to learn a deformation field, which are therefore weakly-supervised. In the following, details of each network in this framework as well as the loss terms wills be explained.

4.2.1 Segmentation network

Some recent works that use spatial transformer networks for image registration utilize image label information to improve the registration accuracy. [Hu et al., 2018c] introduced an auxiliary loss term based on manual labels in their work. To account for potentially missing labels, only one structure was selected at random among those present in each iteration. The labels that represent important structures or keypoints in both images are smoothed using a Gaussian smoothing kernel. In the training, each

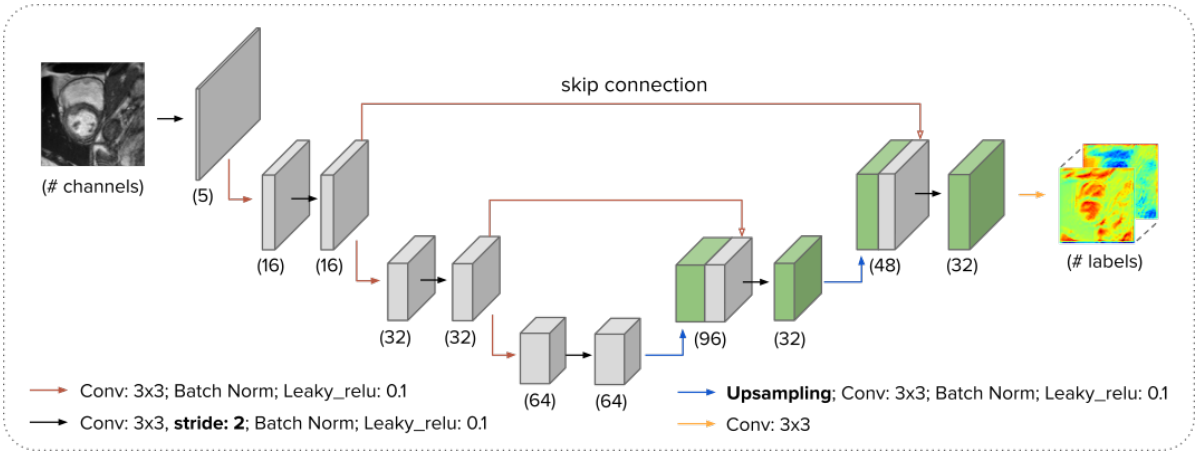


Figure 4.2: Segmentation U-Net architecture. A simple U-Net structure with two skip connections is trained to learn to segment important organs. The segmentation network can be pre-trained with supervision or trained simultaneously with the registration networks in an end-to-end fashion.

label is then selected in each iteration randomly to compute a soft Dice loss between the warped fixed and moving labels. In our framework, the semantic information provided by segmentation is employed explicitly for the feature extraction of both images individually. The feature extraction network is shown in Figure 4.2. The simple U-Net designed for our framework has 11 convolutional layers with the kernel size of 3×3, two skip connections, three downsampling layers and two upsampling layers. The output has half the size of the input image. In most cases, especially where the semantic structures represented with a segmentation are large enough, the smaller size of the output does not affect the accuracy of the network. The number of channels of the network can be adjusted based on the number of labels as well as the complexity of the image data.

The network takes a grayscale image $\mathsf{I} : \mathbf{x} \mapsto \mathbb{R}$, $\mathbf{x} \in \Omega$ and outputs a SoftMax prediction $\mathbf{f} \in [0,1]^{L \times N}$ for N sampled points and for each label $l \in 1, \dots L$. A weighted cross-entropy is computed between the output is and the ground truth labels, which is our semantic loss:

$$\mathcal{L}_{\text{semantic}} = -\sum_{j}^{N} \mathbf{w}\hat{\mathbf{f}}(\mathbf{x}_j) \log \mathbf{f}(\mathbf{x}_j), \tag{4.1}$$

where $\mathbf{w} \in \mathbb{R}^L$ is the label weight vector, $\hat{\mathbf{f}}$ is the ground truth segmentation and $\mathbf{x}_j$ is the j-th sample point. The label weight is computed as the inverse class frequency for each label in the training dataset. The network weights are shared for both fixed and moving images I_f and I_m.

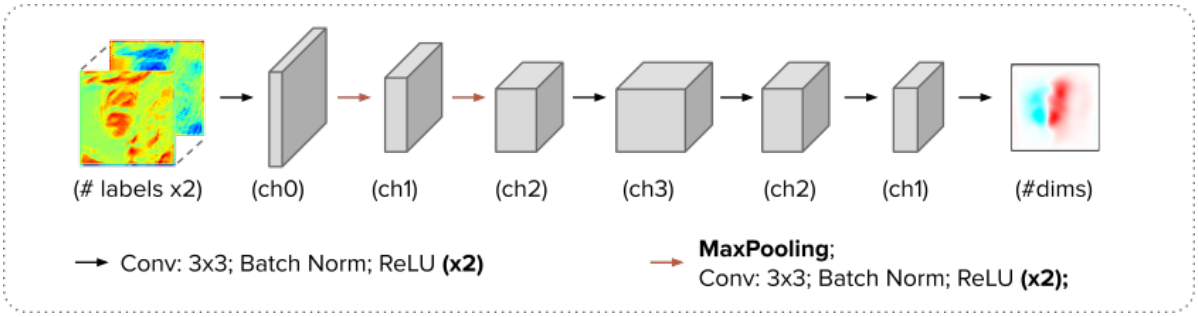


Figure 4.3: Network architecture used for registration. Max-pooling layers are used to reduce the dimensionality of the input and to train the network to generalize better.

4.2.2 Registration network

Given the network outputs $\mathbf{f}_f, \mathbf{f}_m \in [0,1]^{C \times H_s \times W_s}$ of the feature extraction network, the registration network estimates a deformation field based on the semantic information, i.e. the class probability, embedded in the features. The number of channels C is the same as the number of segmentation classes L including the background and H_s and W_s are the size of the output in each spatial dimension and $N = H_s W_s$. The features are concatenated and passed on to the first registration network, which outputs a dense deformation estimation $\mathbf{V}$, which comprises the displacement values at each pixel. The network architecture of the registration network is depicted in Figure 4.3. The registration network consists of 13 convolutional layers with kernel sizes of 3 for all spatial dimensions, each followed by a batch normalization and a ReLU layer. As in [Ha et al., 2020], a max-pooling layer is used after the first two convolutional layers to reduce the dimensionality of the input.

Based on the estimated deformation field $\mathbf{V}$, the deformation loss $\mathcal{L}_{deform}$ using semantic information is computed as:

$$\mathcal{L}_{deform} = \frac{1}{L} \sum_{l=1}^{L} w_l \mid \mathsf{S}_f(l) - \mathcal{T}(\mathsf{V}, \mathsf{S}_m(l)) \mid, \tag{4.2}$$

where w_l denote the label weights and $\mathcal{T}$ is the spatial transformer for the B-spline transformation. The estimated deformation field $\mathbf{V}$ is upsampled to meet the input image size using bilinear (or trilinear in 3D case) interpolation, and an identity grid is added to the generated sampling grid. The moving segmentation image $\mathbf{S}_m$ is then spatially transformed by resampling using the resulting grid.

In addition to the deformation loss, a regularization loss of the estimated deformation field is computed to ensure smooth global transformation. The regularization loss $\mathcal{L}_{regular}$ is computed as:

$$\mathcal{L}_{\text{regular}} = ||\mathsf{V} - \mathsf{V}_{smooth}||_2. \tag{4.3}$$

where V_{smooth} is the locally smoothed deformation field.

Our final loss term $\mathcal{L}$ for semantically-guided deformation estimation is then:

$$\mathcal{L} = \lambda_s \mathcal{L}_{semantic} + \mathcal{L}_{deform} + \lambda_r \mathcal{L}_{regular}\,, \tag{4.4}$$

where λ_s and λ_r are the weight parameters for the semantic and regularization loss terms.

4.3 Experiments

In this section, first the datasets used for our experiment and the preprocessing steps performed for each dataset are described. Next, the implementation details for our framework will be explained. Finally, a short description on the experimental settings will be given.

4.3.1 Datasets and preprocessing

For our experiment, we use the Helen face dataset provided by Smith et al. [2013] and Automated Cardiac Diagnosis Challenge (ACDC) training dataset from the MICCAI Challenge 2017 [Bernard et al., 2018].

The Helen face dataset consists of 2330 samples of 2D face RGB images as well as segmentation labels of face structures (face, eyes, eyebrows, nose, mouth and hair). The images are converted into grayscale and cropped to have the same pixel dimensions of 320×260 using an enlarged face bounding box. We used 2000 images for training and 330 images for test as in the work of Le et al. [2012]. Since inner mouth labels are absent in some samples, we merge three mouth structures (upper lip, lower lip and inner mouth) into a single label. In addition, we exclude the hair label that often has obscure appearance and hence use only 7 labels in total for our experiments. Example images of the preprocessed Helen dataset used in our experiments are shown in Figure 4.4.

The original ACDC training dataset consists of 3D cine-MRI images of 100 patients with evenly distributed subgroups (4 groups with pathological abnormalities, 1 healthy group). The segmentation of left and right ventricles and the myocardium (Figure 4.5) is also provided, along with the diastolic-systolic phase instances. We extract end-systolic and end-diastolic phase from each cine-MRI to generate an image pair to be

Figure 4.4: Example images of the Helen face dataset used in our experiment. Seven structures are marked using different colors: face, left eyebrow, right eyebrow, left eye, right eye, nose and mouth. Original mouth labels (upper lip, lower lip and inner mouth) are combined to a single mouth label.

registered. The extracted images are resampled and cropped into the spatial dimension of $128 \times 128 \times 64$ with pixel spacing of 1.56mm in each dimension. For our experiment, we divide the dataset into four folds, where each fold includes 25 image pairs and three folds are used for training and the remaining fold as test dataset.

4.3.2 Implementation details

The U-Net architecture of our framework is implemented as described in Section 4.2.1 and visualized in Figure 4.2 and is kept the same for the different datasets. The network consists of approximately 200k parameters and two skip connections. To equally account for the different labels regardless of their size, a weighted cross entropy loss is used.

The registration network consists of 13 convolutional layers with two max-pooling layers as depicted in Figure 4.3. The number of channels is selected based on the size of input images, where $ch1 = 2 \cdot ch0$, $ch2 = 2 \cdot ch1$ and $ch3 = 2 \cdot ch2$ and $ch0 = 16$. For the medical dataset, we slightly modify the network architecture by removing the second max-pooling layer, since the image size of the medical dataset is smaller than the face dataset. As illustrated in Figure 4.1, we use two registration networks with the same network architecture successively to build a two-step warping framework.

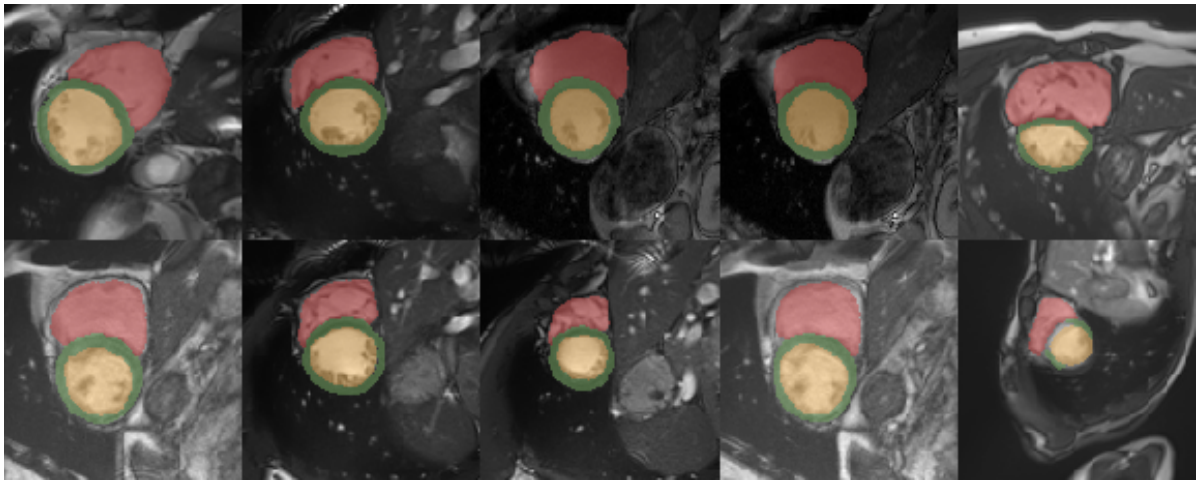


Figure 4.5: Example image slices of ACDC dataset used in our experiment. Three structures are marked using different colors: ■ right ventricle, ■ myocardium and ■ left ventricle.

The outputs of each registration network have smaller spatial dimension (i.e. smaller transform parametrization) than the input features (via max-pooling), which we then upsample into the feature map size and apply three cardinal B-spline smoothing steps using average pooling layers with stride 1. The kernel sizes of these average pooling layers are selected based on the image size, where we chose 19 for the output of the first registration network applied on the face dataset. For the output of the second registration network, we chose a smaller kernel size of 11. The reason for this choice is to enable the first network to capture large and coarse deformation, whereas the second network refines the alignment of smaller structures. We choose smaller kernel sizes (5 and 3 for the outputs of the first and second registration network respectively) for the medical dataset, considering the smaller image size. The kernel sizes are chosen to be approximately 5% and 3% of the largest spatial dimension of each input image, and have been determined empirically.

For the computation of the regularization loss, the smoothed deformation field V_{smooth} was generated by applying average pooling layers twice on the output deformation field of the second registration network. The kernel size of these average pooling layers are chosen to be 5 and 3 for the face and medical dataset, respectively.

The training was performed for 300 epochs for both dataset and the training batch size was chosen to be 20 (for face dataset) and 5 (for medical dataset). The whole framework is optimized in end-to-end manner using Adam optimizer with the learning rate of 0.001 and the momentum of 0.97 for both dataset. We empirically determined the weights of

loss terms, where we use $\lambda_r = 0.001$ and $\lambda_s = 1.0$ for regularization and semantic loss, respectively.

4.3.3 Ablation study

To evaluate the importance of each component of our framework, we perform an ablation study. For the ablation study, we use the face dataset, since the dataset has a large number of training samples and exhibits larger image variations compared to the medical datasets. In the ablation study we experiment with internal modifications such as:

- training using single or multistep registration network,
- training with or without the semantic loss and
- the choice of regularization parameters.

The first two experiments are performed using a pre-trained segmentation network.

4.4 Results and Discussion

Experiments are designed to evaluate the effect of each component of the framework. In the ablation study, we evaluate the effect of using semantic-guidance (4.4.2) and the importance of using multistep registration network (4.4.1). In addition to the ablation study, we also perform an extensive experiment to evaluate label bias (4.4.4). Finally, the result of our network is compared with other recent registration methods.

4.4.1 Single vs. multistep registration networks

While the architectures of registration networks are unchanged, for the single step registration network we use more number of parameters, so that the total number of parameters for compared registration network configurations are similar (approximately 2.3 million parameters). The number of parameters are modified by changing the number of channels of the convolutional layers. We hypothesize that using two smaller networks instead of a single network with a large number of parameters will deliver a better registration results. For a better comparison of the registration part, we use a pre-trained feature network in this experiment. The registration results of single- and two-step registration networks are shown in Figure 4.6. As we hypothesized, the two-step network (*pre-trained unet, two-step*) shown in orange dashed line outperforms the single network (*pre-trained unet, single*) shown in green dashed line (lower line indicates better result).

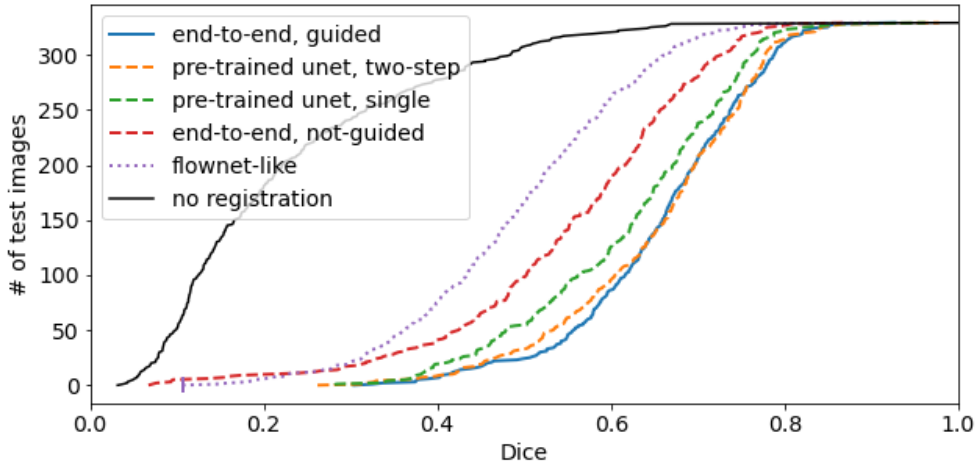


Figure 4.6: Sorted Dice scores (averaged across face structures) for all test images for various network configurations (lower line is better). Graphic is from Ha et al. [2020].

4.4.2 Semantically guided deformation estimation

When training feature and registration networks jointly in an end-to-end manner, we hypothesize that the semantic loss can improve registration accuracy by giving an advantage of training a feature network with a soft constraint. With the end-to-end training, registration networks can benefit from more generic image information such as edges. The results are shown in Figure 4.6 (*end-to-end, guided*: blue solid line and *end-to-end, not-guided*: red dashed line) and Table 4.2 (*ours (without guidance)* and *ours*). The Dice score of our method using semantic guidance is improved by 11.5% compared to the results without guidance.

4.4.3 Regularization

To ensure the smoothness of the estimated deformation field and avoid singularities that leads to implausible deformation, we use regularization term as described in Equation 4.3. The output deformation fields from the registration networks are smoothed using average pooling layers. The kernel size of the average pooling layers influences the smoothness of the final deformation fields, and it is important to consider the size of the structures to be registered. While large structures can benefit from a large kernel size of the pooling layers, smaller structures might not be correctly transformed, and the deformation fields might lose details. Dice values of our network using the regularization loss for each structure is shown in Figure 4.7. Approximately 10% of the small

structures such as eyebrows and eyes are not registered at all, which might be influenced by the occlusion of such structures by hair. Occlusion in the image leads to singularities in the deformation field (higher rates of negative Jacobians), which is penalized by the regularization loss.

An additional experiment using a diffusion regularization resulted in a better accuracy both in the Dice values and the mean contour distances (Table 4.2: *ours (diffusion regularizer)*). The standard deviation of the Jacobians and the rates of negative Jacobians have much higher values compared to the results using regularization loss based on smoothed deformation fields. With more singularity points allowed in the estimated deformation field, the structures that might be partially occluded by hair could be deformed better, resulting in a higher accuracy.

4.4.4 Label-bias

The sorted Dice scores for different structures are shown in Figure 4.7. When all labels are seen during training, the Dice scores are much better for larger structures such as face and nose. However, for smaller structures such as eyes and eyebrows, there are some cases, where these structures are not registered at all (outliers shown as circles). One of the possible reasons for this besides the label size can be the occlusion, since the eyes and eyebrows are covered by hair or glasses in some images. Another reason can be the large variation in the initial position of the smaller structures compared to the face and nose, both of which are located mostly at the center of the image.

In Figure 4.8, the Dice score of each structure is shown in the box plot. In each graph, results from models trained with different labels are shown in different colors. The mean Dice is shown as the dotted line of each box plot, and the red lines indicate the median value. Especially to evaluate influence of the face label which contains the rest of the smaller labels inside, we train the network with only face, all other labels without face and face with each one of the remaining structures. The holes on the face label formed by removing smaller structures are filled in for this experiment, ensuring that the small structures within the face is also considered as parts of the face. Since the face dataset comprises segmentations of small structures that are located within a large face segmentation, the deformation within the face label is influenced by these auxiliary information of smaller labels. If only the face label is observed during training (Figure 4.8: *only face*), it is unlikely that the smaller structures are aligned appropriately, since the deformation in a homogeneous region within face segmentation cannot be adjusted properly.

The best mean Dice score of all structures is achieved, when the face segmentation is not observed during training (Figure 4.8: Mean - *w/o face*). In this case, the small

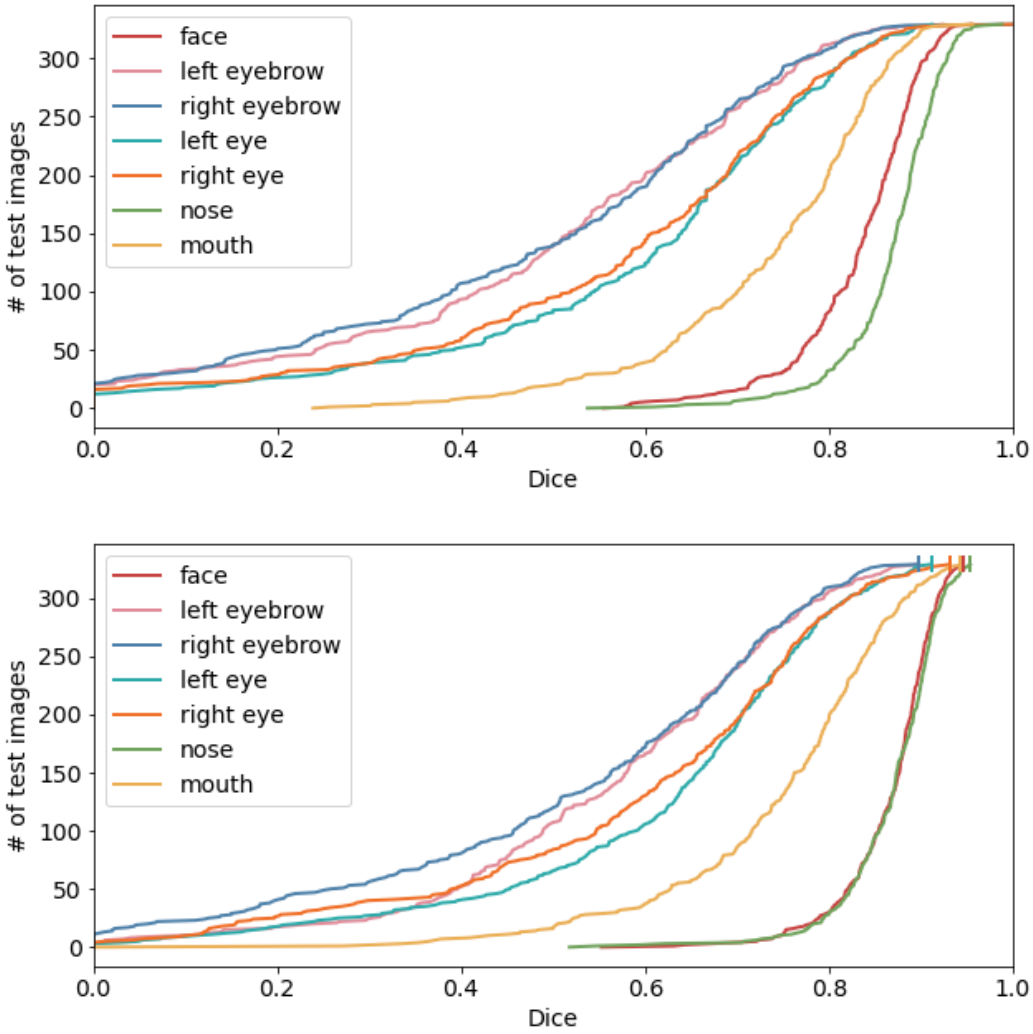


Figure 4.7: Sorted Dice scores for different face structures. The result using regularization loss with smoothed deformation field (top) and the result using diffusion regularization (bottom) is shown. The alignment of eyes and eyebrows is challenging due to occlusions caused by hair in many images (top). However, using diffusion regularization, these structures are better registered (bottom). Graph (top) is from Ha et al. [2020].

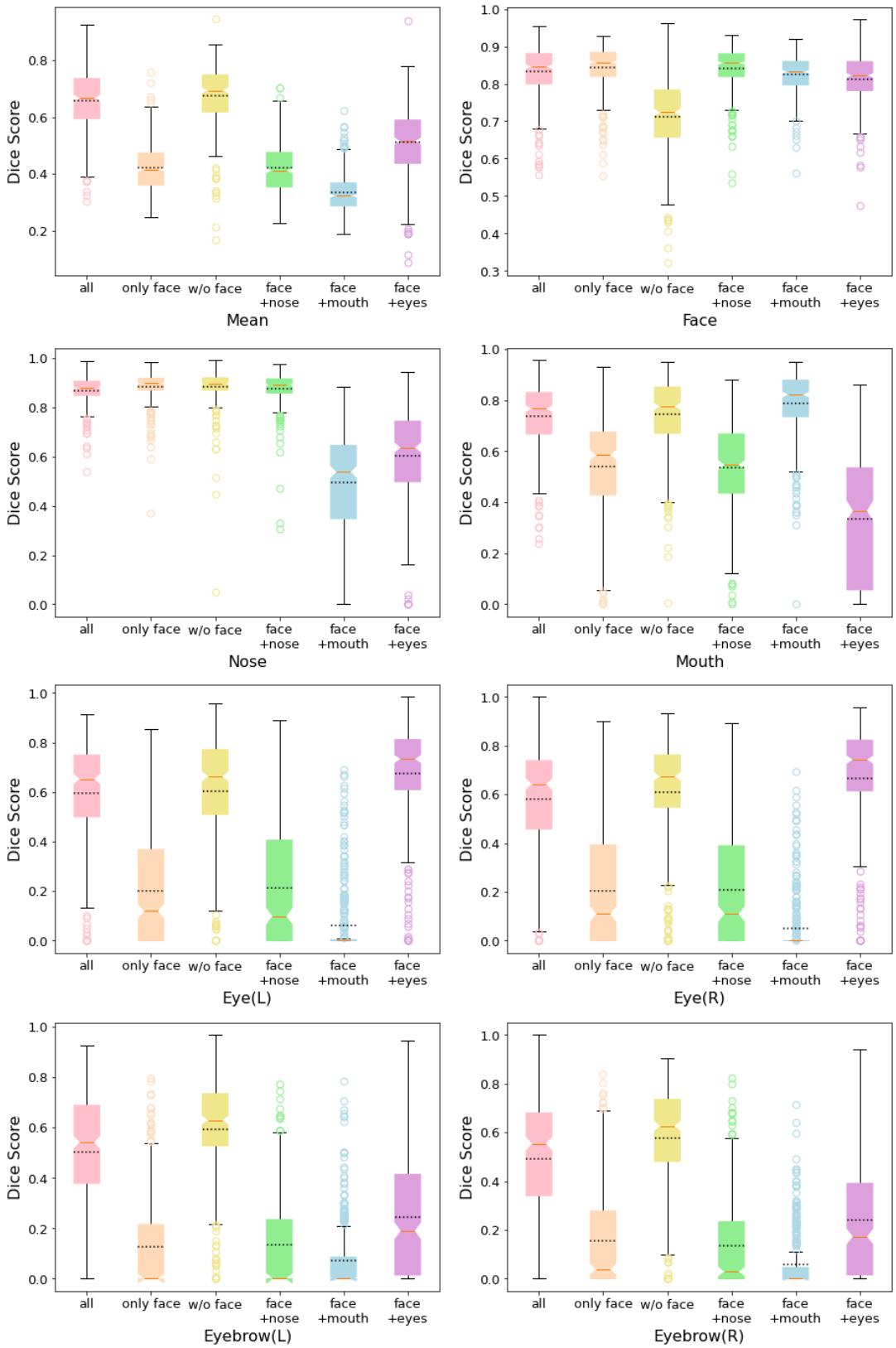


Figure 4.8: Comparison of Dice scores of each label trained with different subsets of labels.

structures can be aligned more freely without being influenced by the face segmentation. As can be expected, most structures cannot be well aligned, if their segmentation is not seen in the training. However, the Dice score was not too low, if the observed structure has strong spatial correlation with the labels observed during training (e.g. nose-eyes).

In addition to the imbalance of registration accuracy influenced by the size and variations of the initial location of the structure, one of the possible shortcomings of the semantic guidance with segmentations (or other labels) is the imbalance in registration accuracy between the structures with and without label. Since the networks in our framework are optimized based on label registration accuracy (i.e. mean squared distance between fixed and warped moving labels), deformation plausibility of the structures without a label cannot be considered in the weight updates of the networks and the registration of these structures can only depend on the nearby structures with a label. To evaluate the influence of observed segmentation on the overall registration accuracy, we train the proposed framework using 1) all segmentation available, 2) only half of randomly selected segmentation and 3) just one segmentation and compare the results. As shown in Figure 4.9, accuracy drops when the number of labels observed during training are reduced. However, the Dice scores of the structures, whose labels are not observed during training (Figure 4.9: *Not observed*) have also improved compared to the initial values (Figure 4.9: *Not observed (before reg)*). This might be because of the spatial correlation to the other structures observed during training. To achieve appropriate accuracy for the proposed method, labels of the most relevant structures should exist, and it might be favorable to consider physical properties of the structures to label.

4.4.5 Comparison with other state-of-the-art methods

Dice score, contour distance, the standard deviation of Jacobian determinant and mean rate of negative Jacobian determinants (number of negative pixels/total number of pixels) of different methods and our method with different settings are compared in Table 4.2. The different methods compared in this experiment were implemented using the same loss function as in the original methods, except for FlowNet. We use the original FlowNet [Ilg et al., 2017] and the pre-trained model, however to constitute an easier registration task, we experiment with the downsampled images, which are pre-aligned using an affine transformation (Table 4.2: *FlowNet w/ smaller images*). We also compare a new variant of FlowNet, which is combined with our proposed B-spline parametrization (Table 4.2: *B-spline FlowNet*). This yields a significant improvement compared to the original FlowNet, however it still has more than 10% points lower accuracy than our proposed method (Table 4.2: *ours*).

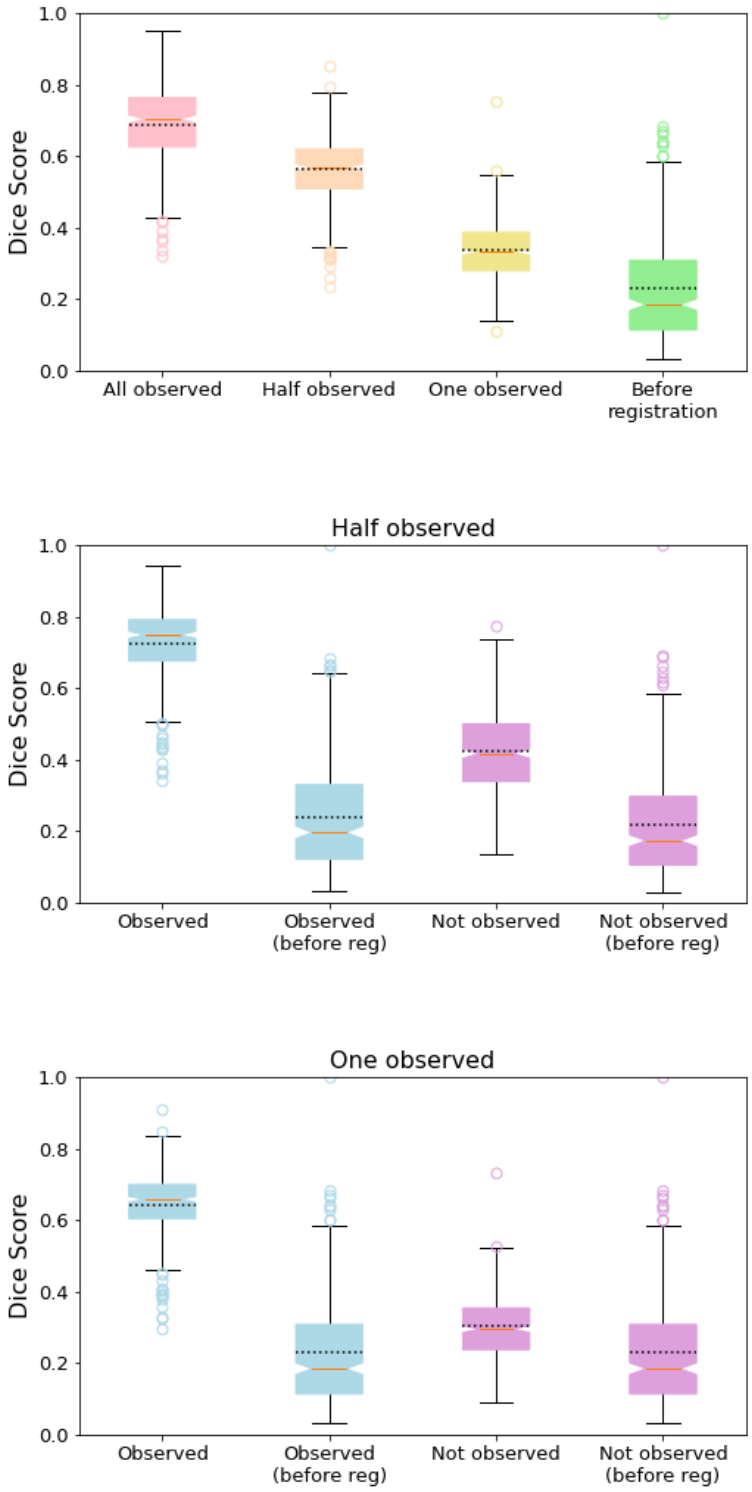


Figure 4.9: Comparison of mean test Dice scores using networks trained with different subset of labels.

Table 4.1: Dice scores, contour distances (Contour Dist.) in pixel, the standard deviation of Jacobian determinant (Jacob. Std.), mean number of negative Jacobian determinants (Jacob. Negatives) of estimated deformation fields using different methods (for the Helen dataset) are shown. For B-Spline FlowNet, ours (without guidance, similar to Hu et al. [2018b])) and ours (shared RegNet weights, similar to [Qin et al., 2018]), we have implemented our own version of the method from each reference. Explicit shape regression (ESR, [Cao et al., 2014]) requires in addition corresponding manual landmarks for training and for the dense warp field coherent point drift algorithm (CPD, [Myronenko and Song, 2010]) is used. The experiment using pre-trained FlowNet was performed without fine-tuning and using downsampled images (Ilg et al. [2017]), which we first affine transformed based on the face landmarks. Approximated inference time (Inf. Time) is also given in seconds per image.

Method	Dice (%)	Contour Dist.(px)	Jacob. Std.	Jacob. Negatives	Inf. Time (s/img)
no registration	23.0	15.55	-	-	-
ESR + CPD	65.6	**1.96**	0.154	-	-
FlowNet w/ smaller images	30.6	12.83	-	-	-
B-Spline FlowNet	49.4	5.96	0.579	0.03124	≈0.007
ours (without guidance)	55.5	5.41	0.257	0.00062	≈0.007
ours (shared RegNet weights)	60.4	5.02	0.269	0.00061	≈0.007
ours	66.0	4.01	0.285	0.00106	≈0.007
ours (diffusions regularizer)	**68.7**	3.67	0.455	0.01459	≈0.007

We implement our method without the semantic guidance, which is comparable to the method proposed by Hu et al. [2018c] (Table 4.2: *ours (without guidance)*). For this experiment, a two-step registration framework is used and the only difference to our proposed method was the absence of the semantic loss. To compare the effect of using a two-step registration against recursive training of registration networks, we implemented our own version of the method that is similar to the method proposed in Qin et al. [2018] (Table 4.2: *ours (shared RegNet weights)*). Keeping the other parts of our framework the same, we only modify the registration networks setting, i.e. used shared weights for registration networks. As summarized in Table 4.2, the Dice score can be improved by 10% using the semantic guidance and the use of a two-step registration framework is more beneficial than using two recurrent networks for registration.

Table 4.2: Dice scores, contour distances (Contour Dist.) in pixel, the standard deviation of Jacobian determinant (Jacob. Std.), mean number of negative Jacobian determinants (Jacob. Negatives) of estimated deformation fields using different methods (for the Helen dataset) are shown. For B-Spline FlowNet, ours (without guidance, similar to Hu et al. [2018b])) and ours (shared RegNet weights, similar to [Qin et al., 2018]), we have implemented our own version of the method from each reference. Explicit shape regression (ESR, [Cao et al., 2014]) requires in addition corresponding manual landmarks for training and for the dense warp field coherent point drift algorithm (CPD, [Myronenko and Song, 2010]) is used. The experiment using pre-trained FlowNet was performed without fine-tuning and using downsampled images (Ilg et al. [2017]), which we first affine transformed based on the face landmarks. Approximated inference time (Inf. Time) is also given in seconds per image.

Method	Dice (%)	Contour Dist.(px)	Jacob. Std.	Jacob. Negatives	Inf. Time (s/img)
no registration	23.0	15.55	-	-	-
ESR + CPD	65.6	**1.96**	0.154	-	-
FlowNet w/ smaller images	30.6	12.83	-	-	-
B-Spline FlowNet	49.4	5.96	0.579	0.03124	≈0.007
ours (without guidance)	55.5	5.41	0.257	0.00062	≈0.007
ours (shared RegNet weights)	60.4	5.02	0.269	0.00061	≈0.007
ours (single RegNet)	62.3	3.93	0.276	0.000896	≈0.007
ours	**68.7**	3.67	0.455	0.01459	≈0.007

Finally, we compare our weakly supervised learning based method with a strongly supervised landmark model. Explicit face regression proposed by Cao et al. [2014] is an example approach in this category that predicts landmarks for all test images. To generate a dense deformation field from the estimated landmarks, we use the coherent point drift algorithm [Myronenko and Song, 2010] (Table 4.2: *ESR + CPD*). Landmark annotations with accurate point-to-point correspondences are however much harder to generate than annotating segmentation labels, and it is particularly difficult to find and define meaningful landmarks in many applications such as for 3D medical images. For this reason, we consider this approach as an upper bound for employing supervision with segmentation only. Still, our method achieves comparably high Dice scores with only slightly more complex transformations, which is reflected by higher Jacobian de-

Figure 4.10: Example results of our approach for the Helen dataset. Top) moving images with the corresponding ground truth labels, (middle) fixed images with the warped ground truth labels of moving images, (bottom) warped moving images. Image from Ha et al. [2020].

terminants. Example registration results using our method is shown in Figure 4.10 for qualitative evaluation.

4.4.6 Medical dataset

In Table 4.3, the Dice score from intra-patient cardiac motion registration results using different unsupervised registration methods are compared for 3D ACDC dataset [Bernard et al., 2018]. The Dice scores of compared methods are reported by Krebs et al. [2018], where left ventricle blood pool (LV-BP) corresponds to L.V. and LV-Myo to Myocard. The compared approaches employed classical optimization-based method [Lorenzi et al., 2013] or unsupervised learning strategies [Balakrishnan et al., 2019; Krebs et al., 2018].

Each pair of images represent the registration of end-systolic and end-diastolic phases of each cardiac cycle, which establishes the problem of large motion and often with strong image artefacts as shown in Figure 4.5. Generalization of the model for such a dataset might be difficult due to large variations in image appearance and contrast across subjects. Classical optimization-based methods and unsupervised learning methods account for intensity levels within one patient to bypass this problem. Our framework

Table 4.3: Mean Dice scores (%) of different methods for medical cardiac dataset averaged across 30 test subjects by labelled structures (R.V.: right ventricle, L.V.: left ventricle). Table from [Ha et al., 2020].

	R.V.	L.V.	Myocard	**Mean**
Unregistered	65.1	66.0	52.5	61.2
LCC-Demons [Lorenzi et al., 2013]	70.6	77.6	73.0	73.7
VoxelMorph [Balakrishnan et al., 2018]	68.1	74.3	71.6	71.3
Krebs et al. [Krebs et al., 2018]	68.4	75.6	74.0	72.6
Ours	**77.4**	**82.5**	**73.4**	**77.8**

can also cope with those problems, where the deformation field is only estimated and optimized based on the semantic information given by labels. The quantitative results shown in Table 4.3 show that our method achieves comparable accuracy to the state-of-the-art approaches compared. Example registration results are shown in Figure 4.11 for qualitative evaluation.

4.5 Summary

In this chapter, a learning based image registration framework using a U-Net and a encoder-like registration network is presented, which is designed to align a pair of images with a large initial misalignment. We train networks in an end-to-end manner using three loss terms to incorporate semantic information, alignment of relevant structures and regularization of the estimated deformation field. Through extensive experiments of the proposed framework on a 2D face dataset, the importance of each module is evaluated and appropriate hyperparameters are determined. The application on a 3D medical dataset also shows a comparable accuracy to state-of-the-art machine learning based registration approaches such as *Label-Reg* [Hu et al., 2018c] and *VoxelMorph* [Balakrishnan et al., 2019].

Although our experiment was limited to mono-modal image registration, registration of multi-modal images can also be considered with an appropriate modification of the feature networks (e.g. using separate networks for each image or not sharing weights for the first few layers of the U-Net), since the network weights are optimized based on semantic information given by the segmentation labels.

One of the limitations of the proposed framework is the label bias, since the framework relies on the semantic information given during the training and therefore requires a careful selection of structures to be labeled. Arrangement of such annotations with

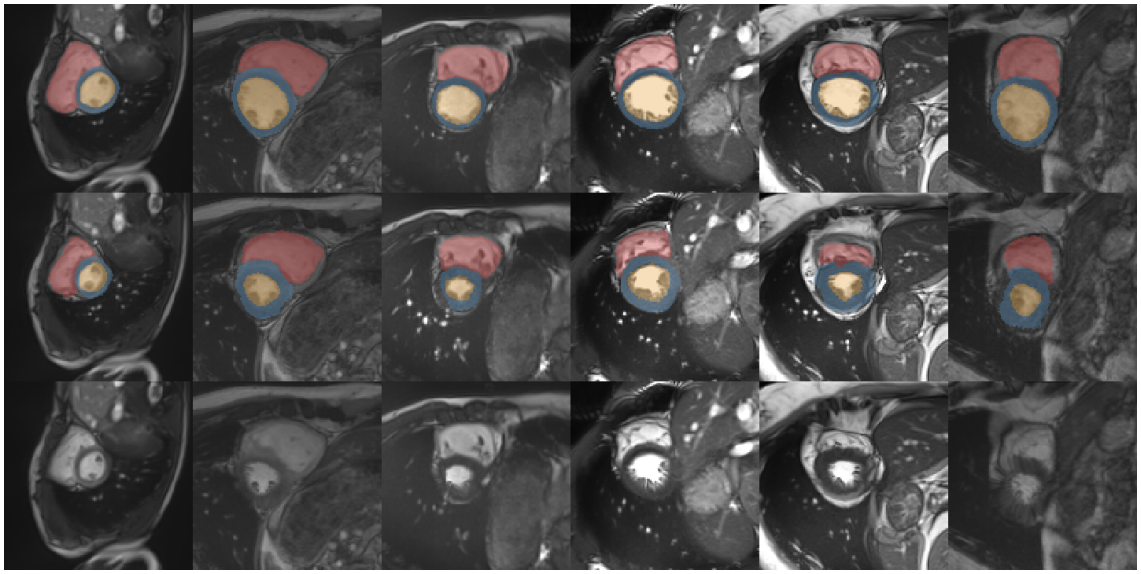


Figure 4.11: Example results (example slices) of our approach for intra-patient registration. (Top) end-systolic slices with the corresponding ground truth labels, (middle) end-diastolic slices with warped ground truth labels of end-systolic slices, (bottom) warped end-systolic slices. Image from [Ha et al., 2020].

appropriate quality might be costly, since manual segmentation is time-consuming and requires skilled experts and quality. On the other hand, adequateness of the segmentation cannot be guaranteed with automated segmentation algorithms. A possible solution to this problem is to use an unsupervised approach, which will be discussed in the next chapter.

Chapter 5

Discrete multi-modal registration for image-guided surgery using deep feature learning and instance optimization

As mentioned in the last chapter, arranging ground truth deformation fields or annotations with appropriate quality for auxiliary information is rather difficult for medical images. To avoid this problem, there have been research on unsupervised image registration approaches using machine (or deep) learning. However, it is still challenging for such approaches to achieve the accuracy and the robustness comparable to state-of-the-art classical image registration approaches. In Ha and Heinrich [2019], published in Lecture Notes in Computer Science, we compared different building blocks of learning based on discrete registration and proposed a new attention module to estimate information content of a grid point. However, learning attention weights in order to determine informative regions that are used to find an optimal transformation did not work as well as we expected.

In this chapter, we extend the previous work and present an image registration approach that combines a self-supervised deep learning model for feature extraction with a conventional discrete image registration approach based on a block-matching algorithm. The efficient block-matching implementation presented in chapter 3 is used to speed up the computation of mapping correspondences. Regularization is performed using a novel instance optimization algorithm, which optimizes estimation of displacement vectors based on local displacement probabilities and global transformation conformity. The learned model for feature extraction is able to extract comparable features from different imaging modalities. In addition, we distill the trained feature network using network pruning for reduction of model complexity and computation time. The presented approach is submitted to

the international open access peer-reviewed journal *Computer Methods and Programs in Biomedicine (CMPB)* [Ha and Heinrich, 2021].

5.1 Introduction

In many clinical image-guided applications, different imaging modalities are used before and during treatment. For example, an imaging modality such as MRI-FLAIR is usually used in image-guided brain surgery to obtain anatomical and pathological information in planning phase, whereas US is used to acquire intraoperative information [Sastry et al., 2017]. In order to monitor (and correct) brain shift and to enable intraoperative navigation, preoperative and intraoperative information obtained by different imaging modalities have to be fused. For this purpose, a multi-modal image registration method is required. In this chapter, we introduce a novel approach for multi-modal image registration, which combines learned modality-agnostic self-supervised features and a post-fitting process that optimizes the registration by incorporating global conformity of the estimated transformation.

In our previous work, we have proposed an end-to-end registration framework that uses attention network and heatmap loss [Ha and Heinrich, 2019]. Although it was shown that the features learned by CNN networks will be as good as using hand-crafted features, finding informative regions using attention network did not work as we expected. Therefore, instead of training an end-to-end network, we adopt an instance optimization algorithm to optimize local displacement probabilities computed based on feature dissimilarities as well as global conformity of the estimated displacements. In addition, the best trained feature network is compressed using network distillation (network pruning) for efficient computation.

5.2 Related Works

For the CuRIOUS challenge in 2018, from total 8 participants the only deep learning algorithm participated [Zhong et al., 2018] has shown a comparable result to the other conventional methods that participated in the challenge evaluated on the training dataset. However, the evaluation on the test dataset did not reach the accuracy of the other methods, which indicates that this algorithm was over-fitted to the training dataset and was not able to generalize well on the test dataset. The conventional approaches that have shown high accuracy on registration of the CuRIOUS challenge dataset use linear correlation of linear combination LC^2 [Fuerst et al., 2014], self-similarity descriptors [Heinrich, 2018a], and attribute matching and mutual-saliency approach [Machado et al.,

2018]. These approaches achieve mean TREs of approximately 2 mm on the test dataset, with the computation time ranging from 20 to 450 seconds, whereas the learning-based approach took only 1.8 seconds on a CPU. In general, the learning-based approaches are still suffering from poor accuracy compared to the conventional registration approaches. However, their capabilities of fast computation can provide a substantial advantage for registration of medical images, especially for image-guided interventions, where realtime image analysis is of interest.

With the increasing number of researches using deep learning, many approaches for deep learning based medical image registration are also proposed in recent years. Many unsupervised approaches adopt a loss based on classical similarity measures such as normalized cross-correlation [de Vos et al., 2017; Yoo et al., 2017; Li and Fan, 2017] in combination with a spatial transformer network [Jaderberg et al., 2015; Yoo et al., 2017; Balakrishnan et al., 2018]. However, such deep learning based approaches have shown only moderate success in image registration of larger deformations. In these works, images are pre-aligned using a rigid transformation [Li and Fan, 2017] or the evaluations are performed on an intra-patient registration task with small variations. The problems of limited alignment accuracy in image pairs with large initial deformations are most likely due to the limited capture range. When the input images are concatenated from the beginning, it might be difficult for a network to learn contextual features relevant for large motion registration. One way to improve outcomes in particular for multi-modal image registration is to accommodate expert labels for supervision. Some recent works utilize labels for CT/MRI registration tasks with simple U-Net based network architectures [Hering et al., 2019a; Hu et al., 2018c], two-step registration with semantic-guidance Ha et al. [2020], generative adversarial networks [Tanner et al., 2018; Xu et al., 2020], shape-encoders [Blendowski et al., 2020a] and multivariate mixture models [Luo and Zhuang, 2020]. An obvious drawback of these approaches is that expert labels are not always available, and the quality of registration outcomes depends upon the quality of the used labels.

5.3 Method

In previous chapters, we have presented a classic discrete registration approach that utilizes a prior knowledge of deformation provided by a motion model and a deep learning approach that utilizes auxiliary semantic information in an end-to-end learning framework. As mentioned in the previous chapter, medical datasets with ground truth deformation fields are difficult to find, and it is costly and time-consuming to arrange segmentation, although it is somewhat simpler to acquire than ground truth deformation fields. In the following, we present an unsupervised image registration approach,

which registers images solely based on information provided by the image itself without incorporating any auxiliary information. We use a self-supervised feature network for feature extraction and then utilize the extracted features to perform a discrete registration between MRI and US images.

To address the challenging task of 3D MRI to 3D US brain registration, we implemented and compared the influence of different components of an end-to-end learning-based framework such as attention module and heatmap loss for large deformations in our previous work [Ha and Heinrich, 2019]. Unfortunately, the registration accuracy of the evaluated end-to-end framework was not comparable to other state-of-the-art approaches, and we observed no advantage of using the attention module and heatmap loss in the experiment. Presumably, training of an attention network to learn appropriate weights in order to sort out unreliable correspondences based only on the TRE loss or heatmap loss is not sufficient, since these rely only on local information. For this reason, we give up on employing such modules for end-to-end learning-based methods and instead propose an elegant optimization algorithm as a post-fitting step. The proposed optimization algorithm simultaneously considers the regularization of estimated displacement fields by penalizing deviation from the global transformation and local displacement probabilities based on the dissimilarity cost map.

An overview of our proposed method is shown in Figure 4.1. Given a fixed image $I_{\mathcal{F}} : \Omega \to \mathbb{R}$ and a moving image $I_{\mathcal{M}} : \Omega' \to \mathbb{R}$ with the image field of view $\Omega, \Omega' \in \mathbb{N}^{H \times W \times D}$, we aim to predict rigid or affine transformation parameters $\theta \in \mathbb{R}^{3\times4}$ to optimally align the two images. First, to enable direct comparison between images from different imaging modalities with non-linear statistical relations, an appropriate feature descriptor that can describe image structures independent of local contrast should be trained. We train a CNN network to extract structural information from input images under the supervision of a state-of-the-art multi-modal feature descriptor, the MIND [Heinrich et al., 2012a] with SSC [Heinrich et al., 2013b] (details on MIND/SSC descriptor in section 2.1.1). With the learned features, further fine-tuning is possible and is also usually faster to compute than handcrafted feature descriptors.

The feature network is trained with random image patches of both preoperative MRI and intraoperative US images using shared weights. The output of the feature network is a feature map describing each pixel with a vector of size $n_{feat} = 48$, which corresponds to the number of channels of the network output. From the output feature maps, N_{cp} control points $\mathbf{x}_c \in \mathbb{R}^3$ are defined on a regular grid. Then, the feature maps $\mathbf{f}_{\mathcal{F}}, \mathbf{f}_{\mathcal{M}} \in \mathbb{R}^{n_{feat} \times h \times w \times d}$ are compared at the control points $\mathbf{x}_c \in \mathbb{R}^3$ and then dissimilarity cost maps are generated based on the differences between the features. For each dissimilarity cost map generated for a control point $\mathbf{C}(\mathbf{x}_c) \in \mathbb{R}^{r \times r \times r}$, the feature of the control point $\mathbf{x}_c$ in the fixed image is compared with the feature of the control points within the

capture range r (times the stride) in the moving image $\mathbf{y}_i$ ($i \in [0, r^3]$). The displacement probabilities of each control point can be calculated by using min-convolutions [Heinrich, 2019] on the corresponding dissimilarity map. Then simply by selecting the local minima of the dissimilarity cost determined by the $argmin$ operation, the initial displacement vectors of all control points $\mathbf{V}_0 \in \mathbb{R}^{3 \times N_{cp}}$ can be determined.

From the initial estimation of the displacement vectors, we can determine the transformation parameters by performing the least squares fitting. However, the parameters directly estimated from the initial displacement vectors are suboptimal, since no relation between control points are considered for this estimation, and it may contain numerous outliers that hamper the optimal fitting to the global transformation. We employ a post-fitting step to tackle this problem, which iteratively optimizes a joint energy function of the locally sampled matching cost (local displacement probability) and global regularization cost of the displacement vectors. During the optimization, the displacement vectors for grid points are updated based on the differentiable trilinear sampling of local displacement probabilities and a penalty that encourages global linear transformation conformity. The final displacement vectors $\mathbf{V}$ is then used to estimate optimal rigid or affine global transformation parameters using a weighted least squares fitting method.

5.3.1 Feature network

During training, a CNN network is trained based on random image patches with the patch size N_p for each dimension simultaneously for both images using shared network weights. A mini-batch consisting of random patches of all training samples is fed into the network as input.

The feature network consists of four convolutional layers, each with a kernel size of 3 except for the first convolutional layer, which has a kernel size of 5. Each convolutional layer is followed by a batch normalization layer and a ReLU, whereas a sigmoidal function is followed after the last convolutional layer. The first convolutional layer also includes dilation of 2, which results in size reduction of the network output. A detailed description of our feature network is given in Table 5.1.

The network is trained using self-supervision of the state-of-the-art handcrafted feature descriptor MIND/SSC [Heinrich et al., 2013b], which extracts modality-independent neighborhood descriptors based on the self-similarity context. Using the MIND/SSC descriptor with a four-neighborhood, we obtain intensity and modality invariant feature vectors for each voxel with a length of 48 (describing four neighboring voxels each with 12 values), which determines the number of output channels of our feature network. The network is optimized based on the SmoothL1Loss computed between the network

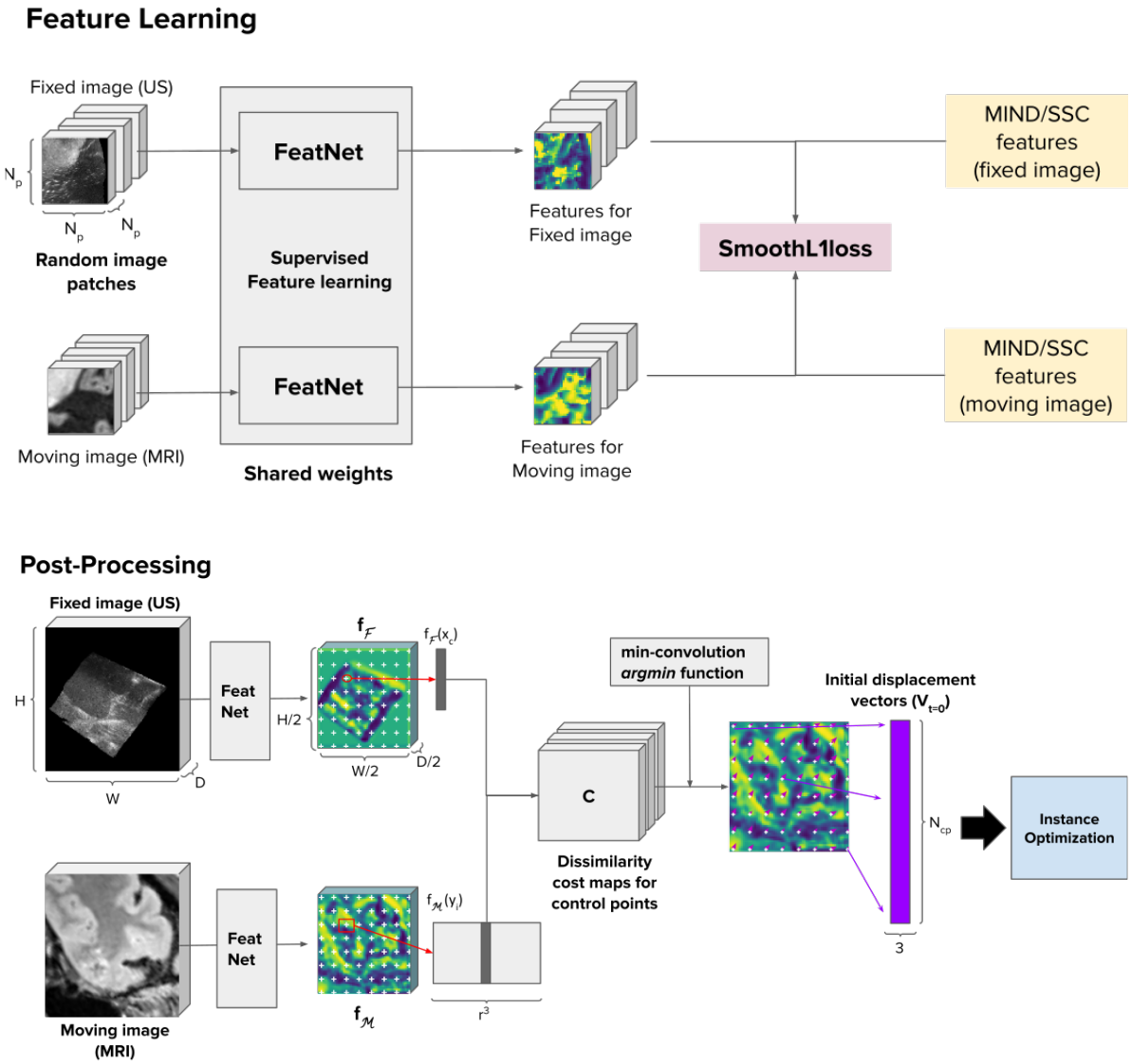


Figure 5.1: An overview of the proposed method. The feature learning is performed previous to the actual estimation of the transformation parameters θ. The registration is performed in the post-processing step using an instance optimization algorithm.

Table 5.1: The network architecture of the feature extraction network. *inch* and *outch* denote number of input and output channels, *kernel* is the kernel size, and *s*, *p*, and *d* denote stride, padding, and dilation respectively.

	inch	outch	kernel	s	p	d
conv3d	1	32	5	1	4	2
batchnorm3d			-			
ReLU			-			
conv3d	32	64	3	2	1	1
batchnorm3d			-			
ReLU			-			
conv3d	64	128	3	1	1	1
batchnorm3d			-			
ReLU			-			
conv3d	128	48	3	1	1	1
sigmoid			-			
# parameters	447,088					

output and the ground truth MIND/SSC features. The SmoothL1Loss calculates the sum of absolute element-wise differences between the inputs and is defined as:

$$L_{smoothL1}(\mathbf{f}_a, \mathbf{f}_b) = \begin{cases} \frac{1}{n}\sum_i 0.5 \cdot (\mathbf{f}_a(i) - \mathbf{f}_b(i))^2/\beta, & if \quad |\mathbf{f}_a(i) - \mathbf{f}_b(i)| < \beta \\ \frac{1}{n}\sum_i |\mathbf{f}_a(i) - \mathbf{f}_b(i)| - 0.5 \cdot \beta, & else \end{cases} \tag{5.1}$$

where $\mathbf{f}_a, \mathbf{f}_b$ are the feature maps, i is the index of the element and β is a threshold usually set to 1. If the absolute difference between two elements is smaller than the threshold $\beta = 1$, the loss is equivalent to mean squared error (MSE) and if β is set to 0, the term is equivalent to L1 loss.

5.3.2 Computation of dissimilarity costs

Based on the feature maps extracted by the feature network, a tensor of the dissimilarity cost map $\mathbf{C} \in \mathbb{R}^{N_{cp} \times r \times r \times r}$ is generated. For every N_g-th voxel located on a regular grid, a dissimilarity cost map is computed. For each control point $\mathbf{x}_c$, the SSD between the feature descriptor of a corresponding image voxel in fixed image $\mathbf{f}_{\mathcal{F}}(\mathbf{x}_c) \in \mathbb{R}^{n_{feat}}$ and the feature descriptors of the image voxels within the capture range r of moving image $\mathbf{f}_{\mathcal{M}}(\mathbf{y}_i)_{i=\{0,\dots,r^3\}}$ are computed. As a result, we obtain dissimilarity cost maps for all control points $\mathbf{C} \in \mathbb{R}^{N_{cp} \times r \times r \times r}$, which is then smoothed using two average pooling layers

with the kernel size of 3 and the stride of 1 in spatial dimensions. Finally, a min-convolution is approximated using a max pooling layer and an average pooling layer with the kernel size of 3 and the stride of 1 in displacement dimensions [Heinrich, 2019]. The initial displacement vectors $\mathbf{V}_0 \in \mathbb{R}^{N_{cp} \times 3}$ for the control point set can be determined using an *argmin* operation on the resulting displacement cost $\mathbf{C}$.

5.3.3 Efficient probabilistic instance optimization

Selecting the most probable displacement based on a mathematical *argmax*/*argmin* operation is the trivial choice for most of the previous approaches, which is also how we determine the initial displacement vectors for control points. The selected displacements can be used to find a suitable global linear transformation with either 6 parameters (rigid transformation) or 12 parameters (affine transformation) using a weighted least square optimization algorithm. However, in this case, an optimal displacement for each control point is determined individually without considering their relation to the neighboring control points. It might result in inaccurate estimation of a global transformation, since the estimated displacement vectors may contain numerous outliers that are largely inconsistent with the other displacement vectors. This problem can be solved by incorporating an appropriate regularization loss into the optimization process. In this work, we introduce a probabilistic instance optimization-based approach that combines local displacement probability with a regularization cost based on a global transformation. This extends the concept presented in Heinrich [2019] for linear transformations.

The main idea is to use an efficient Adam based optimization [Kingma and Ba, 2014] to minimize a joint energy function consisting of matching cost and regularization cost, based on the pre-computed densely sampled dissimilarity costs $\mathbf{C}$. A simple network consisting of parameters $\mathbf{W}_{params} \in \mathbb{R}^{N_{cp} \times 3}$ is generated and initialized with the initial displacement vectors $\mathbf{V}_0$ determined by using the *argmin* operation based on the dissimilarity costs. The network parameters, i.e. the estimation of displacements for all control points, can be considered and updated simultaneously during the iterative optimization process via loss terms. Graphical description of instance optimization process is depicted in Figure 5.2.

The probabilities of subvoxel displacements can be estimated using a trilinear interpolation, and they can be sampled for control points using the network parameters $\mathbf{V}_t$ in each iteration t. Using the sampled displacement probabilities $\mathbf{C}_{sampled,t}$, the local probability loss $\mathcal{L}_{local}$ is computed as:

$$\mathcal{L}_{local} = -\frac{1}{N_{cp}} \sum_{i}^{N_{cp}} \Lambda \mathbf{C}_{sampled} \tag{5.2}$$

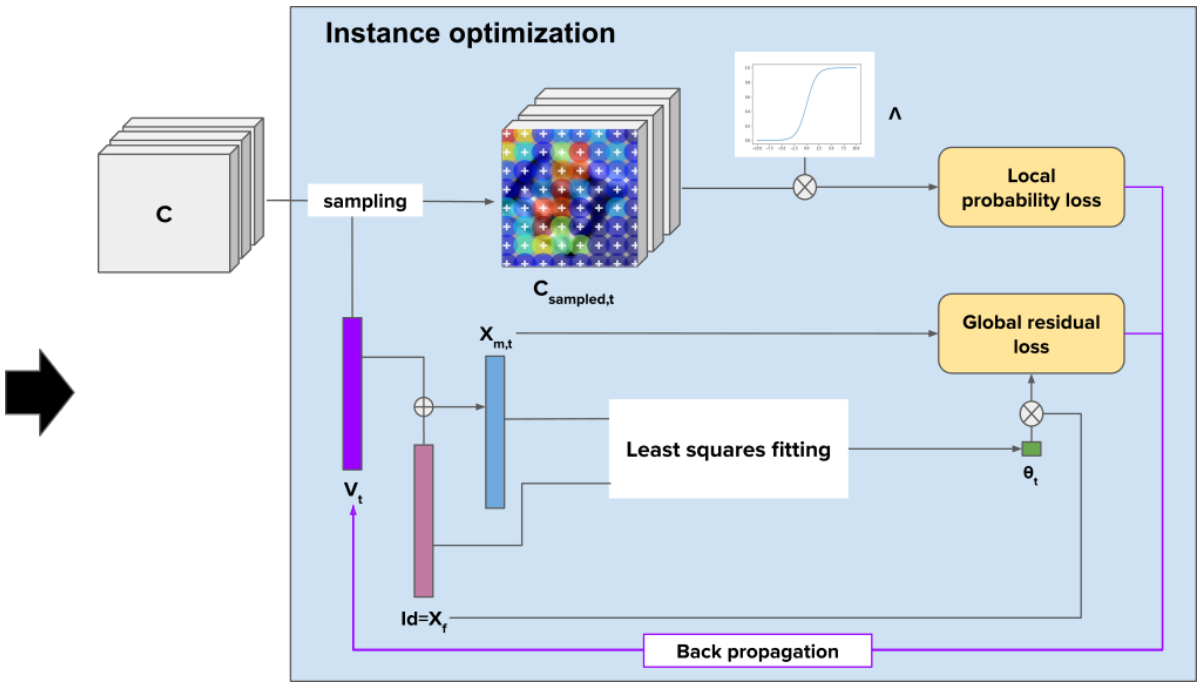


Figure 5.2: Graphical description of instance optimization. The network parameter V_t, initialized with initial displacement vectors V_0, is updated in each iteration based on local probability loss and global residual loss. The local displacement probability at t-th iteration $C_{sampled,t}$ is sampled from the dissimilarity cost maps C using the network parameters v_t and the control points with low probability are removed using the weights (Λ). The output transformation parameter at t-th iteration θ_t is used to compute global transformation conformity, i.e. global residual loss.

where Λ is the weights for sampled points, which remove the sampled points with small displacement probability (Figure 5.2 weights). Together with the local displacement probability loss computed using the sampled displacement probabilities, a regularization loss (or global residual loss) is incorporated for updates of the network parameters.

To compute the regularization loss, the MSE between the globally warped point set for fixed image and the point set for moving image are calculated. The point sets are generated based on the identity grid **Id**. The point set for fixed image $\mathbf{X}_{\mathcal{F}}$ is equivalent to the identity grid **Id** and the point set for moving image $\mathbf{X}_{\mathcal{M}}$ is calculated as the sum of **Id** and the estimated displacement vectors $\mathbf{V}_t$. Global transformation parameters θ_t are computed using a weighted linear least squares fitting method of the generated point sets, where the same weights Λ applied for local displacement loss are used. With the resulting global transformation parameters, the regularization loss $\mathcal{L}_{reg}$ is computed as:

$$\mathcal{L}_{reg} = ||\mathcal{T}_{\theta_t}(\mathbf{X}_{\mathcal{F}}) - \mathbf{X}_{\mathcal{M}}||^2 \tag{5.3}$$

where $\mathcal{T}_{\theta_t}(\mathbf{X}_\mathcal{F})$ denotes the globally transformed point set for fixed image with transformation parameters θ. This regularization loss term of instance optimization is designed to penalize deviations of the displacements at control points from the global transformation. An optimal consensus of locally high displacement probabilities and a globally linear transformation is sought iteratively in this way.

5.3.4 Network pruning

Network pruning is an approach to reduce the size of a train neural network by systematically removing parameters of it. The contribution of each parameter of a trained network to the final output varies. Based on their contribution, we can remove the parameters with lower importance to reduce the size of the network. There are two ways to prune the network; unstructured and structured. With unstructured pruning, i.e. weight pruning that sets individual values of the weight matrices to zero, we can obtain a sparse network. However, the network cannot be compressed easily in this case and requires special software or hardware for the compression [Liu et al., 2019]. On the contrary, structured pruning works at the level of channel or layers, which makes it easier to compress the network without special software or hardware. In this work, we perform a channel-level structured pruning as suggested in Li et al. [2016], where unimportant channels can be sorted out from the network using scale factors and sparsity induced penalty. A schematic illustration of the performed pruning process is visualized in Figure 5.3.

Given a function family $f(x;\cdot)$ of an untrained network with a training input x, the trained network can be defined as $f(x;W)$ with specific network parameters W. For network pruning, the trained network $f(x;W)$ is first fine-tuned using a new loss term:

$$L_{fine-tuning} = \sum_{(x,y)} l(f(x;W),y) + \lambda \sum_{\gamma\in\Gamma} g(\gamma) \tag{5.4}$$

where x and y denote the network input and ground truth respectively, $f(x;W)$ is the network output and l is a training loss function. The first term corresponds to the normal training loss used to train the network, whereas the second term is newly added for fine-tuning. A weight parameter λ establishes a proper balance between two terms. The function $g(\cdot)$ is a sparsity-induced penalty and γ is the scaling factor for each channel. After the fine-tuning, the input and output end of a channel with low importance will also have near-zero values. Thereby, the problem of accuracy loss after pruning can be reduced. After removing $k\%$ of the network parameters from the fine-tuned network $f(x;W')$, the size of the network can be reduced, and we obtain a new smaller network $f(x;\tilde{W}')$ with $|\tilde{W}'| < |W'| = |W|$. In this work, we performed network pruning on our feature network. During fine-tuning, $g(\cdot)$ is chosen to be the L1-norm of the weights of

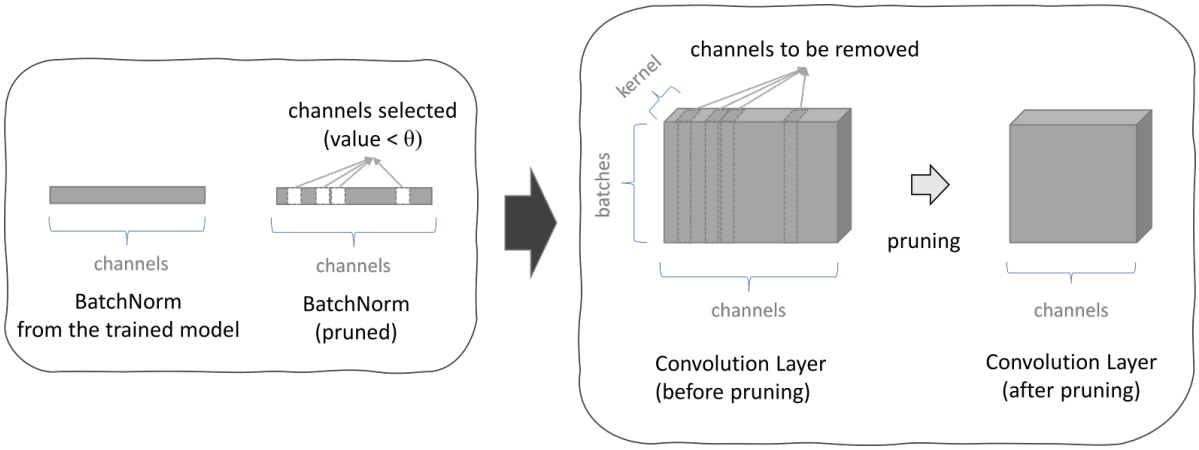


Figure 5.3: Schematic illustration of the pruning process. In the first iteration, the channels to be removed are selected by comparing the values of the BatchNorm layer output with an appropriate threshold value θ and then in the second iteration, the selected channels are removed from the convolution layers and BatchNorm layers, reducing the size of the network.

the batch normalization layers in the feature network as proposed in [Li et al., 2016] and λ is set to 0.004.

5.4 Experiments

The CuRIOUS Challenge demonstrates the problem of multi-modal image registration and provides an open framework to evaluate algorithms. The images of the CuRIOUS Challenge are from an image-guided brain resection surgery, where MRI FLAIR and T1 images are taken preoperatively and US images after craniotomy are taken intraoperatively, i.e. in actual operation after craniotomy and after brain resection. The MRI image taken in the planning phase and the US image taken directly after opening the skull should be aligned before the resection of the tumor to correct the brain shift. We evaluate our approach on the dataset from the CuRIOUS Challenge, where we align preoperative FLAIR MRI images on US images obtained after craniotomy before resection. Example image slices are shown in Figure 5.4.

The training dataset of the CuRIOUS challenge consists of 22 cases for registration of preoperative MRI images and an intraoperative US image. There are 4 scans for each case, a T1-weighted MRI, a FLAIR scan, and an intraoperative US scan after craniotomy and after the brain resection. In our experiment, we register FLIAR images to US scans taken directly after craniotomy. Experiments are performed on 19 training cases of the challenge dataset, which include cases 1-8, 12-19, 21, 23, and 24. The images are resam-

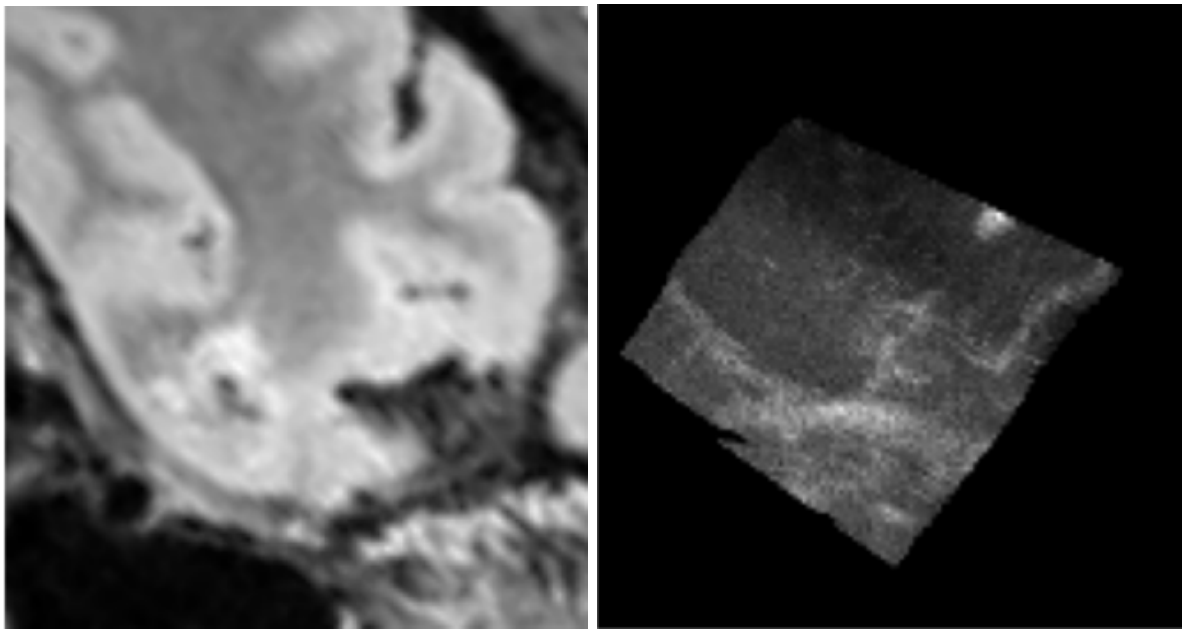


Figure 5.4: Example image slices of the CuRIOUS dataset used in our experiments. A preoperative MRI image of the brain is shown on the left, and an intraoperative US image from the same patient before tumor resection is shown on the right.

pled to isotropic voxel sizes of 0.5 mm^3 in each image dimension, and the annotation landmarks are only used in inference for an evaluation purpose. The landmarks are converted into 3D spheres as described in Heinrich [2018a] for the evaluation. A five-fold cross-validation is performed, with each training fold comprises seven or eight cases containing two-thirds of the difficult cases with a large initial misalignment. Evaluation results are computed by averaging each evaluation case in all folds.

The size of the input random patches for the feature network is selected to $N_p = 64$ voxels for all spatial dimensions. The feature network is trained using SmoothL1Loss with the threshold $\beta = 1$, which means the SmoothL1Loss behaves like the MSE function, when the range of errors is below 1. To avoid using only the mean squared error function for loss computation, we scale the feature network output and the ground truth features by multiplying both with the scale factor of 4. In this way, the value range of the network output and ground truth features can be adjusted from [0,1] to [0,4] and the SmoothL1Loss will not behave like MSE for the values larger than 1. The learning rate of the feature network is set to 0.002 and training was performed for 500 epochs. The best model with the lowest training loss is stored and used for inference. A regular grid is defined for every seventh voxel in each spatial dimension of the image ($N_g = 7$) on which the control points are set for computation of dissimilarity costs. The default cap-

ture range (or search region) for each control point covers 35% of each image dimension, which contains 11^3 discrete displacement steps ($dw = 11$).

For the parameter network generated for the instance optimization, the spatial dimensions of the 5D parameter tensor are calculated as $h_g = \lfloor \frac{H}{N_g} \rfloor$, $w_g = \lfloor \frac{W}{N_g} \rfloor$, and $d_g = \lfloor \frac{D}{N_g} \rfloor$. The network parameters are initialized with the displacement vectors $\mathbf{V}_0$. The network is optimized using an Adam optimizer for 500 epochs with a learning rate of 0.02, if not stated otherwise. We use a sigmoidal weighting function for computation of both loss terms to reduce the influence of potential outliers with large dissimilarity costs. The weights are determined by mapping the sorted ranks of sampled dissimilarity cost of all control points to a range between 0 and 1.

Evaluation is performed based on TREs computed using the ground truth landmarks. Using the estimated rigid/affine transformation matrix, we transform the segmentation spheres of the moving image generated from the ground truth landmarks. The mean TRE is computed between the center-of-mass of the fixed and transformed moving spheres.

5.4.1 Ablation study

For evaluation of the proposed instance optimization algorithm, we compare the rigid/affine transformation parameters, estimated using:

1. only top half of the control points determined directly based on their dissimilarity costs (cf. [Heinrich, 2018a])

2. displacement vectors determined by the instance optimization algorithm, which is optimized based only on the displacement probability loss

3. displacement vectors determined by the instance optimization algorithm, which is optimized based both on the displacement probability loss and the global transformation conformity (regularization loss).

With the first configuration, we remove the displacement vectors from the initial estimation that have low displacement probabilities (or high dissimilarity cost) to optimize the linear fitting. In this case, the displacement probability of each control point is considered independently. In the second configuration, the displacement vectors are optimized simultaneously. However, global conformity of displacement vectors are not reflected during optimization. By incorporating the regularization loss as in Equation 5.3, the deviation of the updated local transformations (displacement vectors) from the estimated global linear transformation can be penalized and the influence of the local transformations with less conformity can be reduced for the weighted linear least squares fitting.

One of the most important hyperparameters that might influence the accuracy of the prediction is the capture range. Since the maximal displacement range that can be captured by the method is controlled by the capture range, it is important to choose an appropriate value for this parameter. To determine the best value for capture range empirically, we compare different values for capture range.

5.5 Results and Discussion

5.5.1 Ablation study

The mean TREs of three configurations are shown in Table 5.2 for each test case. Even when the displacement vectors with smaller displacement probabilities are removed, the estimation result of a weighted linear least squares method of the initial displacement vectors does not work well in most cases. The mean TRE (of all cases) of rigid transformation using the first configuration (Table 5.2 -1) shows no change compared to the mean TRE before the registration. With the affine transformation, the mean TRE improvement is infinitesimal. In more than half of the cases, TREs after the registration are increased compared to the TREs before registration. With the instance optimization without regularization loss (Table 5.2-2), a rigid transformation improves the accuracy, whereas an affine transformation resulted in an even higher mean TRE than before the registration. With our proposed method (Table 5.2-3), the displacement vectors are updated to better fit in the global linear transformation while having high local displacement probabilities. The mean TRE is improved significantly ($p < 0.05$) for both rigid and affine transformation, showing the importance of the additional regularization loss.

In Figure 5.5, the mean TRE of test cases using different capture range is presented. The values indicate the proportion to the image size, which is the same for each image dimension. Using a small capture range, the alignment error of cases with a large deformation cannot be corrected. However, using a too large capture range also influence the accuracy of the prediction negatively. This might be due to the higher possibility of finding the wrong minimum with a larger capture range for cases with small initial misalignment. The best result is achieved with the capture range of 0.35, which is used as the default capture range throughout the following experiments.

The most time-consuming part of our proposed method in inference time is the instance optimization. Although we believe a good convergence of the optimization can be achieved with 500 iterations, this leads to a long computation time. However, the computation time is as essential for intraoperative registration tasks as the accuracy. To find the best trade-off between accuracy and computation time, we experiment with different

Table 5.2: Comparison of TREs (in mm) between the different registration settings. The transformation parameters are determined by (1) an *argmin* operation, displacement vectors of the top half of control points are determined directly based on the dissimilarity costs [Heinrich, 2018a] (2), after the instance optimization considering only displacement probabilities (2) and after the instance optimization using both displacement probabilities and global transformation conformity (3). TREs of rigid transformation and affine transformation for five-fold cross-validation are shown. Table from Ha and Heinrich [2021].

Case	Before Registration	rigid transformation			affine transformation		
		1	2	3	1	2	3
1	1.86	2.54	3.66	**1.67**	2.09	3.73	**2.02**
2	5.75	5.47	5.98	**3.21**	5.86	7.04	**3.34**
3	9.63	8.94	**6.47**	9.17	9.92	**8.97**	9.27
4	2.98	3.77	**2.72**	3.07	2.61	4.49	**1.74**
5	12.20	14.80	5.11	**1.70**	12.34	8.00	**1.99**
6	3.34	4.20	4.58	**1.76**	3.65	5.77	**1.96**
7	1.88	2.82	3.11	**2.06**	**2.27**	4.98	3.25
8	2.65	4.01	3.35	**2.27**	4.34	5.14	**2.71**
12	19.76	12.97	8.00	**1.44**	14.05	10.05	**1.75**
13	4.71	6.49	5.35	**3.52**	5.20	8.30	**2.73**
14	3.03	4.50	3.47	**1.64**	3.50	4.67	**2.11**
15	3.37	**2.56**	5.39	2.98	**3.39**	6.11	3.69
16	3.41	3.54	2.69	**1.73**	3.25	4.75	**1.34**
17	6.41	3.95	3.95	**1.80**	5.63	5.63	**1.49**
18	3.66	2.81	2.16	**1.60**	3.25	5.01	**1.41**
19	3.16	3.18	3.15	**1.95**	3.04	5.39	**2.69**
21	4.46	7.27	5.79	**2.69**	5.88	8.95	**5.51**
23	7.05	4.93	2.26	**1.47**	4.68	6.17	**1.44**
24	1.13	1.78	2.35	**1.42**	2.02	2.63	**1.33**
Mean	5.29	5.29	4.19	**2.48**	5.10	6.09	**2.72**

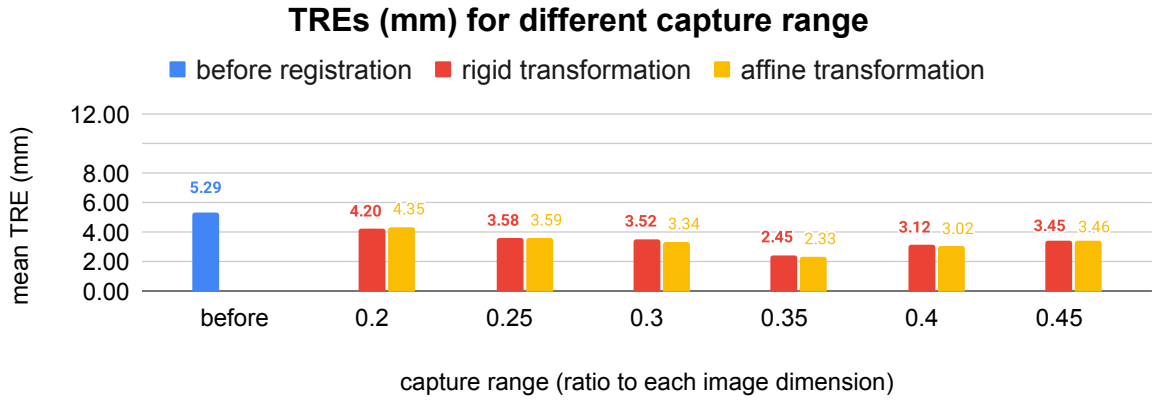


Figure 5.5: Comparison of mean TREs (mm) of different capture range. The number of displacement steps is kept to 11. Graphic from Ha and Heinrich [2021].

numbers of iterations for our instance optimization. Some important experiment results are summarized in Table 5.3. We use different learning rates for different numbers of iterations to find the best convergence rate for each configuration. On the NVIDIA Quadro P6000 GPU, the computation time of our instance optimization algorithm increases by approximately 1 second for 100 iterations.

5.5.2 Instance optimization

Comparison between the direct estimation of global transformation parameters from the displacement vectors determined by an *argmin* operation (1) and estimation after our proposed instance optimization (2 and 3) is presented in Table 5.2. Even when we filter out the control points with low local displacement probabilities for a better the least squares fitting, there is nearly no improvement in TREs compared to the TREs before the registration. When the displacement vectors are determined only based on the individual displacement probability of each control point (2), no coherence between control points can be accounted into the update procedure. It is also difficult to filter out the outliers properly only based on the weighting, that it might still strongly distorts the global transformation estimation, leading to poor registration results.

The proposed instance optimization improves the accuracy by 1.1 mm if rigid transformation is estimated. However, the accuracy is decreased by approximately 1.0 mm for affine transformation. In Table 5.2-2, we can observe that the TRE values after applying the estimated affine transformation are larger in the cases with smaller initial errors. When conformity with the global linear transformation is considered during the instance optimization via regularization loss, the accuracy improves significantly in both estimations of rigid and affine transformations. This demonstrates the importance of

Table 5.3: TREs (in mm) and computation time (in seconds) after instance optimization using different hyperparameters are shown. For both rigid and affine transformation, 500, 200 and 50 of iterations with learning rates of 0.02, 0.04 and 1.2 are compared respectively. Five-fold cross-validation is performed. Table from Ha and Heinrich [2021].

Case	rigid transformation			affine transformation		
	500	200	50	500	200	50
1	1.66	1.65	**1.61**	2.02	**2.01**	2.14
2	3.23	**2.96**	3.28	3.34	3.24	**3.09**
3	**9.17**	9.30	9.32	9.27	9.38	**8.62**
4	3.08	2.77	**1.89**	1.75	**1.61**	1.95
5	**1.70**	11.52	8.56	**1.99**	11.38	7.04
6	1.75	**1.53**	1.76	1.96	1.84	**1.64**
7	**2.06**	2.11	**2.06**	3.25	3.40	**2.76**
8	**2.23**	2.44	2.35	2.70	2.75	**2.40**
12	1.45	**1.40**	12.03	**1.74**	1.77	7.29
13	3.52	3.57	**3.08**	2.73	2.70	**2.69**
14	1.66	1.59	**1.46**	2.21	**1.87**	1.94
15	2.96	2.79	**2.34**	3.68	3.25	**2.91**
16	1.73	**1.70**	1.72	1.33	**1.30**	**1.30**
17	1.82	**1.80**	1.84	1.50	**1.46**	1.73
18	1.60	1.63	**1.45**	**1.40**	1.44	1.41
19	1.96	**1.87**	2.16	2.70	**2.68**	3.20
21	2.72	**1.74**	1.77	5.50	2.94	**2.05**
23	**1.46**	1.53	2.02	**1.46**	1.47	2.08
24	1.42	**1.33**	1.40	1.34	1.31	**1.30**
mean	**2.48**	2.91	3.27	**2.73**	3.04	3.03
duration (s/img)	6.91	3.18	0.82	6.91	3.07	0.82

the information propagation of displacement likelihoods across control points. Without regularization, it is difficult to determine a good set of displacement vectors, which will help find optimal global transformation parameters using the least squares fitting.

The most relevant experiment result for the different hyperparameters and the numbers of iterations for the instance optimization is summarized in Table 5.3. To find an optimal trade-off between accuracy and computation time, we have performed the instance optimization with a different combination of the numbers of iterations and learning rates. With our default setting (500 iterations and a learning rate of 0.02), it takes approximately 7 seconds to register a single case. For some clinical applications, this might be too slow. It takes approximately a second for 100 iterations on a GPU (NVIDIA Quadro P6000) and to enable realtime computation (i.e. under 1 second), the number of iterations should be restricted to be smaller than 100. We have performed 50 iterations for instance optimization with a learning rate of 1.2 and 200 iterations with a learning rate of 0.04. Although realtime computation can be achieved with 50 iterations, the alignment accuracy is dropped by more than 1 mm. With 200 iterations, we can reduce the influence on registration accuracy, while reducing the computation time to approximately 3 seconds.

5.5.3 Network pruning

We experiment with different values for k to find the best pruning rate and the result is summarized in Table 5.4. The mean TRE (in mm) before the registration, the mean TRE of the best model, the fine-tuned model, and the pruned model (with different pruning rates) are shown for rigid and affine transformation. The number of network parameters, model size (in MB), and computation time (s/img), which includes the time until the initial displacement vectors are computed, are also summarized. The TREs of both rigid and affine transformation increase slightly after the model fine-tuning, most likely due to the rearrangement of weights. The best accuracy is achieved when approximately 20% of the network parameters are pruned ($\approx 137,000$ parameters are removed). The result of the pruned model is even better than the original best model in this case for both rigid and affine transformation. The size of the network is reduced by approximately 0.2 MB for every 10% of the removed network parameters, and the computation time by approximately 0.01 second.

In Table 5.5, the final results using the pruned feature network with different instance optimization parameters are summarized. Although the computation time is only slightly reduced, the accuracy is improved with 500 and 200 iterations, indicating that the generalizability of the network is not damaged by the network pruning. Example images (overlay of US images on top of MR images) of the registration result using

Table 5.4: TREs (in mm), number of network parameters, model size, and mean computation time of a single case during inference are compared for before and after pruning. Five-fold cross-validation is performed. The mean computation time includes feature extraction and computation of initial displacement vectors. For network pruning, the best model trained using full network parameters is sparse-tuned with sparsity loss of the network weights first (weight for sparsity loss = 0.004). Different percentages of lower-ranking channels are then removed from the convolutional layers to reduce the network size. The number of parameters after pruning is given in percentage. Table from Ha and Heinrich [2021].

	Before	Best Model	Fine-tuned	Pruned model (%)			
				88.85	78.47	68.57	60.61
rigid	5.29	2.49	2.64	3.01	**2.46**	3.87	3.65
affine		2.67	3.04	2.96	**2.35**	3.75	3.60
#params (1000)	-	447	447	373	310	250	198
model size (MB)	-	1.72	1.72	1.44	1.20	0.999	0.800
duration (s/img)	-	0.92	0.92	0.89	0.89	0.87	0.86

estimated rigid transformation is presented in Figure 5.6 for a qualitative evaluation. The large initial alignment error is significantly compensated by the rigid transformation, which can be observed by the alignment of edge-like structures in the US and MR images (right column in the example).

5.5.4 Comparison with other state-of-the-art methods

We compare our results with the results reported by Xiao et al. [2019]. In the first section of Table 5.6, the results from the CuRIOUS Challenge in 2018 are summarized. All methods listed, except for FAX [Zhong et al., 2018], are classic registration methods, which require relatively long computation times ranging from 20 seconds to 450 seconds. These approaches might not be appropriate for application in some intraoperative image registrations.

The FAX approach takes only 1.8 seconds of computation time on a CPU, which can be improved when a GPU is used. The registration result on the training dataset is also better than the other compared methods. However, the evaluation result on the test dataset indicates that the proposed method is over-fitted to the training dataset and cannot properly register new image pairs. Our proposed method, both with rigid and

Table 5.5: TREs (in mm) and computation time (seconds/case) of five-fold cross-validation result using (80%) pruned feature network. The result of using 500, 200 and 50 iterations for instance optimization are shown (learning rate of 0.02, 0.04 and 1.2 respectively). Table from Ha and Heinrich [2021].

Case	rigid transformation			affine transformation		
	500	200	50	500	200	50
1	1.81	1.77	**1.66**	**1.85**	1.94	2.08
2	2.85	**2.76**	3.06	2.97	**2.92**	2.99
3	**1.43**	4.14	7.92	**1.47**	3.77	8.18
4	2.95	2.72	**2.23**	1.53	**1.49**	2.29
5	9.68	10.51	**9.24**	8.95	10.44	**8.77**
6	1.90	**1.86**	2.39	1.83	**1.78**	2.18
7	2.23	2.05	**1.83**	2.34	1.81	**1.67**
8	2.44	2.46	**2.38**	2.81	2.83	**2.36**
12	**1.46**	1.47	15.08	**1.76**	1.82	10.67
13	**3.06**	**3.06**	3.50	**2.61**	2.66	2.75
14	1.58	1.63	**1.57**	1.60	**1.57**	1.70
15	3.06	2.83	**2.37**	3.04	2.99	**2.69**
16	1.65	**1.64**	1.76	**1.25**	1.27	1.64
17	**1.70**	**1.70**	2.06	**1.52**	1.54	2.18
18	1.66	1.62	**1.59**	**1.50**	**1.50**	1.65
19	1.95	**1.90**	2.51	2.62	**2.57**	3.59
21	2.06	2.12	**1.96**	1.61	**1.58**	1.72
23	**1.61**	1.74	2.27	**1.60**	**1.60**	3.26
24	1.52	**1.43**	1.48	1.43	1.40	**1.36**
mean	**2.48**	2.60	3.52	**2.33**	2.50	3.35
duration (s/img)	5.57	3.14	0.77	5.46	3.03	0.66

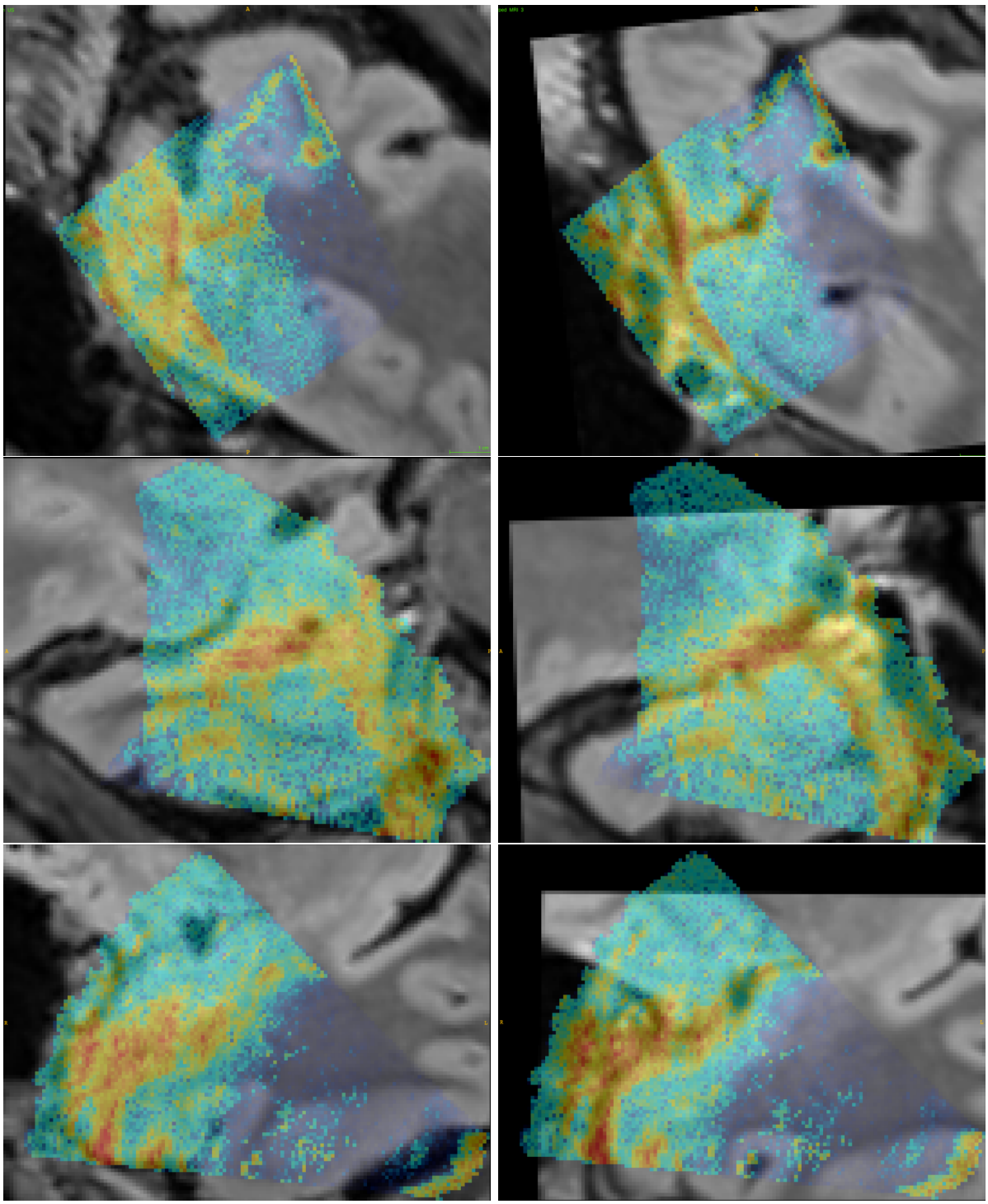


Figure 5.6: Example images of the case 3 registration result. The color overlay of US image slices (in jet) on top of the MRI images (in gray) before registration is shown in the left column. In the second column, the color overlay of US image slices is shown on top of the transformed MRI images, which is aligned better compared to the left column. From the first row, the axial, sagittal, and coronal slices are shown respectively. Images from Ha and Heinrich [2021].

Table 5.6: Mean TREs in mm and computation time of an image pair in seconds from different state-of-the-art approaches. The first section of the table shows the results reported in Xiao et al. [2019]. In the last section of the table are the results from our method. Direct comparison cannot be made between the results in different sections, since the evaluation is performed using a different subset.

method	mTRE (mm)		comp. time
	Training set	*Test set*	
eDRAMMS [Machado et al., 2018]	3.35 ± 1.39	2.18 ± 1.23	450s (CPU)
DeedsSSC [Heinrich, 2018b]	1.67 ± 0.54	1.87 ± 0.93	25s (CPU)
FAX [Zhong et al., 2018]	**1.21 ± 0.55**	5.70 ± 2.93	**1.8s (CPU)**
ImFusion [Wein, 2018]	1.75 ± 0.62	**1.57 ± 0.96**	20s (GPU)
MedICAL [Shams et al., 2018]	4.60 ± 3.40	6.59 ± 2.89	103s (CPU)
NiftyReg [Drobny et al., 2018]	2.90 ± 3.59	3.21 ± 3.57	115s (GPU)
ours (rigid)	2.45 ± 1.84		5.57s (GPU)
ours (affine)	2.33 ± 1.70		5.46s (GPU)

affine transformation, shows comparable results to the compared state-of-the-art classic registration methods with a significantly reduced computation time.

5.6 Summary

In this chapter, we have presented an unsupervised registration approach for the registration of preoperative MRI images and intraoperative US images using modality-agnostic self-supervised deep feature learning and a fast instance optimization as a post-fitting algorithm. To provide suitable CNN features to enable comparison between images from different modalities, a feature network is trained with random image patches with self-supervision by a state-of-the-art handcrafted neighborhood descriptor. Dissimilarity costs are computed for a selected set of control points defined on a regular grid based on the feature map output of the network. Dense displacement probabilities are sampled from the dissimilarity cost map and are used to perform an efficient probabilistic instance optimization to estimate an optimal global transformation for image registration. With exhaustive experiments, we have determined optimal hyperparameters for an appropriate trade-off between accuracy and computation time. In addition, network pruning is performed on the trained feature network for the reduction of model complexity and

computation time. Our proposed method achieves the mean accuracy of approximately 2.50 mm in approximately 3 seconds for the registration of a 3D multi-modal image pair.

Further research can focus on exploring computational gains by learning the update steps of the proposed instance optimization to reduce the number of iterations and thereby further minimize the computation time. It can be accelerated by incorporating the fitting process into an end-to-end learning process, which requires additional development of a deep learning algorithm for a partial differential equation solving [Sirignano and Spiliopoulos, 2018] and the least squares fitting algorithm that can be backpropagated [Wang et al., 2019].

Chapter 6

Conclusion and outlook

In this thesis, three different fast medical image registration approaches for applications in image-guided interventions are proposed. In most (clinical) image-guided interventions, anatomical images with high spatial resolution are acquired for planning of a procedure to accurately locate the target site and to carefully design the procedure for minimization of possible risks during the operation (or treatment) as well as side effects after the procedure. These images can be used to design a statistical model or to train a data-driven deep learning model to provide an important prior knowledge, as well as act as a mapping function for image registration. Image registration is particularly essential for online plan adaptation of image-guided procedures, since the images acquired should be compared continuously during the procedure to adapt the plan. It is also commonly required for frequent setting or adjustment of the plan before starting an actual procedure. Therefore, the aim of the methods presented in this thesis is to utilize those planning images to develop image registration methods, which can align a pair of mono- or multi-modal images ideally in realtime with high accuracy, especially in presence of a large initial misalignment. Although the overarching aims of all three methods developed are shared, the specific focus and application of each method is different and poses complementary challenges.

The first method (Chapter 3) focuses on improving the computation speed of image registration by sparsifying a state-of-the-art conventional regularized block-matching method and is evaluated for a specific application scenario. The second method (Chapter 4) is developed using CNN networks, which can be trained without ground truth displacements. Since inference can already be done in realtime using a deep learning model trained in an end-to-end manner, more focus was on the improvement of the registration accuracy (rather than runtime) for this method. Finally, the third method is developed to predict global transformation parameters in a multi-modal setting based on input images, by combining a deep feature learning CNN network with a conventional matching and novel optimization process and the focus was on both improvement of accuracy and computation time.

For each method, a brief summary of the method and our contribution can be described as the following:

- In Chapter 3, a statistical model based deformable registration method is proposed. The proposed method can perform a realtime deformable image registration for respiratory motion estimation by matching sparse keypoints and reconstructing a dense deformation field from it. The method is evaluated on a scenario of online motion adaptation in MRI- and US-guided interventions, where mono-modal 2D-2D, 2D-3D, and 3D-3D deformable image registrations are performed for estimation of patient-specific respiratory motion. Given an accurate image registration method to provide the ground truth data for patient-specific respiratory motion, the proposed method is able to estimate relatively large respiratory motion for the whole image field-of-view based on sparse keypoints matching result and a statistical motion model generated based on the ground truth deformation fields. A joint optimization of a similarity cost computed using a GPU-accelerated block-matching of sparse keypoints and a regularization cost using the patient-specific motion model is performed using a coupled convex optimization algorithm. Evaluation on different temporal datasets shows that the proposed method can achieve a comparable accuracy to the gold standard method, which usually takes significantly longer time for the same task. Especially, the registration of 2D-3D MRI images achieved the computation time of approximately 3 ms per image pair compared to 60 s of the conventional method, which leaves enough time for dynamic plan adaptation.

- In Chapter 4, an end-to-end weakly-supervised deep learning framework is proposed for deformable image registration. Instead of training CNN networks using ground truth deformation fields, the proposed method trains the networks indirectly using available annotation such as segmentation labels and for an inference of new data using the trained model requires only the raw input images. In particular, we use an explicit loss term to constrain the network training with semantic guidance to enable the network to learn important structural information from the input images. With an extensive ablation study, we have shown that using this explicit loss term for semantic guidance improved the accuracy of prediction. Moreover, we utilize two-step registration networks to further improve the registration accuracy. Using two similarly structured networks in a cascaded manner, smaller deformations which were not always captured by the first registration network can be compensated by the second network. The evaluation on a small 3D cardiac dataset for registration of end-systolic and end-diastolic phases has shown

improved accuracy compared to other state-of-the-art unsupervised deformable image registration approaches.

- In Chapter 5, we propose an unsupervised image registration method, where we combine deep feature learning with a conventional matching method and a novel instance optimization as a post-fitting process. The method is evaluated on a multi-modal dataset for registration of preoperative MRI image and intraoperative US image in image-guided brain surgery. Due to the differences in coordinate system and image field-of-view and the fact that almost no soft tissue deformation is expected for this initial US scans, we restricted our method to predict a global transformation. The CNN network for feature extraction is trained with a self-supervision using a state-of-the-art modality-agnostic feature descriptor that enables comparison between images from different modalities. We compress the trained feature network using network pruning, which have reduced the model size and the inference time, while preserving the accuracy. Using a novel instance optimization method that optimizes global transformation parameters based on a joint energy function of local displacement probability and global transformation conformity, the registration accuracy is improved. In the evaluation on 3D multi-modal image registration data, it has been shown that the proposed method can achieve nearly comparable accuracy compared to other conventional state-of-the-art approaches which requires minimal computation time of 20 s, while being more than 6 times faster.

Although the evaluations were performed on mono-modal image pairs for methods presented in Chapter 3 and Chapter 4, these methods can be extended for multi-modal image registration without significant changes, since the images are described by structure based feature descriptors. For Chapter 4, no specific application scenario was given, however it can be applied for any image-guided interventions, for which enough planning images with structural annotations are available for training. In the following, the other limitations of each method and further extension possibilities are discussed.

- The most significant drawback of the method proposed in Chapter 3 is the need of ground truth deformation fields for generation of a statistical model. Since deformation fields cannot be generated manually, a gold standard method with high accuracy and robustness is required, and the accuracy of the method strongly depends on the accuracy of this gold standard method. One possible solution for this problem is to use deep learning techniques (as discussed in Chapter 4), which enables the training of a learning based model using annotations such as segmentation or landmarks or even for a population model that can be more easily adopted to each patient only based on input images.

- As also discussed in Chapter 4, using segmentation labels to train feature networks inevitably leads to a label bias problem. While the registration accuracy of the structures with a label can be improved using the semantic guidance, the accuracy of the other structures without labels might be less accurate. Another fundamental problem is that it is usually expensive and time-consuming to generate manual annotations, particularly since common medical images are obtained in 3D, and it requires medical experts to produce annotations with an adequate quality, which should be done for each 2D image slice. Using an unsupervised approach can help deal with this problem. However, unsupervised approaches have their own weakness, particularly in the medical domain, because they may struggle in learning to represent highly variable anatomies. To further improve the proposed framework, experiments using different CNN network architectures can be considered. In particular, for an extended application on multi-modal image data, the architecture for the segmentation network can be modified to use different weights for the first few layers (without parameter sharing), which has shown better performance in some recent works [Hering et al., 2019a; Blendowski et al., 2020b].

- The method presented in Chapter 5 combines a deep learning feature network with an instance optimization algorithm. Due to the optimization part, the inference of a new image pair takes still relatively long time for a realtime application. However, as mentioned in the chapter, the proposed method can be extended by substituting some part of the conventional optimization algorithms with learning approaches, enabling differentiable backpropagation. Further research can be performed on expanding the presented method into an end-to-end framework for prediction of transformation parameters, which will improve both the registration accuracy and computation speed. One of the other limitation of the current method is that the prediction is limited to predict global transformation parameters instead of deformable transformation parameters, which might not be suitable for other clinical applications. The extension of the method to predict a deformable transformation model is however possible by for example, performing B-spline interpolation based on the estimated displacement vectors of the control points.

In addition to the limitations mentioned for each method, all three methods presented in this thesis share the problem of using a data-driven model, i.e. data bias (or domain bias). This bias can be alleviated in some degrees, if a large amount of data is available, which is however not always feasible in medical domain. The scarcity of available data may also lead to over-fitting of the model on the training data, resulting in poor inference results for new data. Many recent researches utilize different data augmentation techniques for training of deep learning networks or transfer learning [Raghu et al., 2019] to

deal with this problem. Another possibility to reduce the data bias is the online adaptation of the model, i.e. updating the model continuously with new data. The difficulties of arranging ground truth data of auxiliary structural information can be solved with registration frameworks that are trained in an unsupervised manner.

Another general problem of medical image registration is that unlike segmentation or classification tasks, it is difficult to quantitatively evaluate accuracy of deformable image registration method due to infeasibility of manual annotation of ground truth data. Even in the case, where we use a gold standard method, the evaluation of its accuracy should be performed somehow based on landmarks or segmentations manually annotated on salient structures. This requires again the work of medical experts, which is expensive and time-consuming and will usually only cover parts of the image domain. Some very recent works focus on alleviating this problem using deep learning techniques to help validation of image registration [Eppenhof et al., 2018; Fu et al., 2019].

The use of deep learning techniques for medical image registration enables, in particular, improvement of computation speed via efficient convolutional filters and semantic structures can be distinguished using segmentation networks. If automatic image registration can be trained to work well with scarce image data, the speed of image acquisition time of imaging modalities can also be further improved. However, there is not yet a generally applicable deep learning registration network, analogous to the U-Net for segmentation tasks and it requires integration of a problem-specific a priori model (e.g. for regularization) from classical methods. Although various strategies for deep learning based image registration are explored and the registration accuracy of such approaches are improving, online instance optimization, such as the one introduced in Chapter 3 for motion tracking, might still be important for better accuracy and sufficient computation speed of medical image registration.

Bibliography

Andrade, N., Faria, F. A., and Cappabianco, F. A. M. (2018). A practical review on medical image registration: from rigid to deep learning based approaches. In *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 463–470. IEEE.

Avants, B. B., Epstein, C. L., Grossman, M., and Gee, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical image analysis*, 12(1):26–41.

Balakrishnan, G., Zhao, A., Sabuncu, M. R., Guttag, J., and Dalca, A. V. (2018). An unsupervised learning model for deformable medical image registration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9252–9260.

Balakrishnan, G., Zhao, A., Sabuncu, M. R., Guttag, J., and Dalca, A. V. (2019). Voxelmorph: a learning framework for deformable medical image registration. *IEEE transactions on medical imaging*, 38(8):1788–1800.

Banerjee, J., Klink, C., Peters, E. D., Niessen, W. J., Moelker, A., and van Walsum, T. (2015). Fast and robust 3D ultrasound registration – block and game theoretic matching. *Medical Image Analysis*, 20(1):173 – 183.

Baumgartner, C. F., Kolbitsch, C., McClelland, J. R., Rueckert, D., and King, A. P. (2017). Autoadaptive motion modelling for MR-based respiratory motion estimation. *Medical image analysis*, 35:83–100.

Bay, H., Tuytelaars, T., and Van Gool, L. (2006). Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer.

Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.-A., Cetin, I., Lekadir, K., Camara, O., Ballester, M. A. G., et al. (2018). Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37(11):2514–2525.

Bjerre, T., Crijns, S., af Rosenschöld, P., Aznar, M., Specht, L., Larsen, R., and Keall, P. (2013). Three-dimensional MRI-linac intra-fraction guidance using multiple orthogonal cine-MRI planes. *Physics in Medicine and Biology*, 58(14):4943.

Blendowski, M., Bouteldja, N., and Heinrich, M. P. (2020a). Multimodal 3D medical image registration guided by shape encoder–decoder networks. *International Journal of Computer Assisted Radiology and Surgery*, 15(2):269–276.

Blendowski, M., Hansen, L., and Heinrich, M. P. (2020b). Weakly-supervised learning of multi-modal features for regularised iterative descent in 3d image registration. *Medical Image Analysis*, 67:101822.

Bibliography

Boveiri, H. R., Khayami, R., Javidan, R., and MehdiZadeh, A. R. (2020). Medical image registration using deep neural networks: A comprehensive review. *arXiv preprint arXiv:2002.03401*.

Boye, D., Samei, G., Schmidt, J., Székely, G., and Tanner, C. (2013). Population based modeling of respiratory lung motion and prediction from partial information. In *Medical Imaging 2013: Image Processing*, volume 8669, page 86690U. International Society for Optics and Photonics.

Brix, L., Ringgaard, S., Sørensen, T. S., and Poulsen, P. R. (2014). Three-dimensional liver motion tracking using real-time two-dimensional MRI. *Medical Physics*, 41(4):042302–n/a. 042302.

Brwon, G. and de Vries (2018). White paper: Elekta unity for magnetic resonance radiation therapy (MR/RT). Technical Report 4513 371 1584, Elekta AB.

Calonder, M., Lepetit, V., Strecha, C., and Fua, P. (2010). Brief: Binary robust independent elementary features. In *European conference on computer vision*, pages 778–792. Springer.

Cao, X., Wei, Y., Wen, F., and Sun, J. (2014). Face alignment by explicit shape regression. *International Journal of Computer Vision*, 107(2):177–190.

Cerviño, L. I., Du, J., and Jiang, S. B. (2011). MRI-guided tumor tracking in lung cancer radiotherapy. *Physics in Medicine and Biology*, 56(13):3773.

Chen, M., Lu, W., Chen, Q., Ruchala, K. J., and Olivera, G. H. (2008). A simple fixed-point approach to invert a deformation field. *Medical physics*, 35(1):81–88.

Cheng, J., Tsai, Y.-H., Wang, S., and Yang, M.-H. (2017). Segflow: Joint learning for video object segmentation and optical flow. In *Proceedings of the IEEE international conference on computer vision*, pages 686–695.

Dalca, A. V., Balakrishnan, G., Guttag, J., and Sabuncu, M. R. (2019). Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. *Medical image analysis*, 57:226–236.

De Luca, V., Benz, T., Kondo, S., König, L., Lübke, D., Rothlübbers, S., Somphone, O., Allaire, S., Bell, M. L., Chung, D., et al. (2015). The 2014 liver ultrasound tracking benchmark. *Physics in Medicine & Biology*, 60(14):5571.

De Luca, V., Tanner, C., and Székely, G. (2012). Speeding-up image registration for repetitive motion scenarios. In *IEEE International Symposium on Biomedical Imaging 2012*, pages 1355–1358. IEEE.

De Luca, V., Tschannen, M., Székely, G., and Tanner, C. (2013). A learning-based approach for fast and robust vessel tracking in long ultrasound sequences. In *MICCAI 2013*, volume 8149, page 518. Springer Berlin Heidelberg.

de Senneville, B. D., Hamidi, A. E., and Moonen, C. (2015). A direct PCA-based approach for real-time description of physiological organ deformations. *IEEE Transactions on Medical Imaging*, 34(4):974–982.

de Senneville, B. D., Ries, M., Bartels, L. W., and Moonen, C. T. W. (2012). *Interventional Magnetic Resonance Imaging*, chapter MRI-Guided High-Intensity Focused Ultrasound Sonication of Liver and Kidney, pages 349–366. Springer.

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

de Vos, B. D., Berendsen, F. F., Viergever, M. A., Staring, M., and Išgum, I. (2017). 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*, pages 204–212. Springer.

Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., and Brox, T. (2015). Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766.

Drobny, D., Vercauteren, T., Ourselin, S., and Modat, M. (2018). Registration of MRI and iUS data to compensate brain shift using a symmetric block-matching based approach. In *Simulation, Image Processing, and Ultrasound Systems for Assisted Diagnosis and Navigation*, pages 172–178. Springer.

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

Felzenszwalb, P. F. and Huttenlocher, D. P. (2006). Efficient belief propagation for early vision. *International journal of computer vision*, 70(1):41–54.

Fischer, B. and Modersitzki, J. (2008). Ill-posed medicine—an introduction to image registration. *Inverse Problems*, 24(3):034008.

Fu, Y., Lei, Y., Wang, T., Curran, W. J., Liu, T., and Yang, X. (2020). Deep learning in medical image registration: a review. *Physics in Medicine & Biology*.

Fu, Y., Wu, X., Thomas, A. M., Li, H. H., and Yang, D. (2019). Automatic large quantity landmark pairs detection in 4dct lung images. *Medical physics*, 46(10):4490–4501.

Fuerst, B., Wein, W., Müller, M., and Navab, N. (2014). Automatic ultrasound–MRI registration for neurosurgery using the 2D and 3D LC2 metric. *Medical image analysis*, 18(8):1312–1319.

Gill, S., Li, J., Thomas, J., Bressel, M., Thursky, K., Styles, C., Tai, K., Duchesne, G., and Foroudi, F. (2012). Patient-reported complications from fiducial marker implantation for prostate image-guided radiotherapy. *The British journal of radiology*, 85(1015):1011–1017.

Greenspan, H., Van Ginneken, B., and Summers, R. M. (2016). Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE Transactions on Medical Imaging*, 35(5):1153–1159.

Ha, I. Y., Hansen, L., Wilms, M., and Heinrich, M. P. (2019). Geometric deep learning and heatmap prediction for large deformation registration of abdominal and thoracic CT.

Ha, I. Y. and Heinrich, M. P. (2019). Comparing deep learning strategies and attention mechanisms of discrete registration for multimodal image-guided interventions. In *Large-Scale Annotation of Biomedical Data and Expert Label Synthesis and Hardware Aware Learning for Medical Imaging and Computer Assisted Intervention*, pages 145–151. Springer.

Bibliography

Ha, I. Y. and Heinrich, M. P. (2021). Modality-agnostic self-supervised deep feature learning and fast instance optimisation for multimodal fusion in ultrasound-guided interventions. *Computer Methods and Programs in Biomedicine*, 211:106374.

Ha, I. Y., Wilms, M., Handels, H., and Heinrich, M. P. (2018). Model-based sparse-to-dense image registration for realtime respiratory motion estimation in image-guided interventions. *IEEE Transactions on Biomedical Engineering*, 66(2):302–310.

Ha, I. Y., Wilms, M., and Heinrich, M. (2020). Semantically guided large deformation estimation with deep networks. *Sensors*, 20(5):1392.

Haskins, G., Kruger, U., and Yan, P. (2020). Deep learning in medical image registration: a survey. *Machine Vision and Applications*, 31(1):8.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Heinrich, M. P. (2018a). Intra-operative ultrasound to MRI fusion with a public multimodal discrete registration tool. In *Simulation, Image Processing, and Ultrasound Systems for Assisted Diagnosis and Navigation*, pages 159–164. Springer.

Heinrich, M. P. (2018b). Intra-operative ultrasound to MRI fusion with a public multimodal discrete registration tool. In *Simulation, Image Processing, and Ultrasound Systems for Assisted Diagnosis and Navigation*, pages 159–164. Springer.

Heinrich, M. P. (2019). Closing the gap between deep and conventional image registration using probabilistic dense displacement networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 50–58. Springer.

Heinrich, M. P., Jenkinson, M., Bhushan, M., Matin, T., Gleeson, F. V., Brady, M., and Schnabel, J. A. (2012a). MIND: Modality independent neighbourhood descriptor for multi-modal deformable registration. *Medical image analysis*, 16(7):1423–1435.

Heinrich, M. P., Jenkinson, M., Brady, M., and Schnabel, J. A. (2012b). Globally optimal deformable registration on a minimum spanning tree using dense displacement sampling. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 115–122. Springer.

Heinrich, M. P., Jenkinson, M., Brady, M., and Schnabel, J. A. (2013a). MRF-based deformable registration and ventilation estimation of lung CT. *IEEE transactions on medical imaging*, 32(7):1239–1248.

Heinrich, M. P., Jenkinson, M., Papież, B. W., Brady, M., and Schnabel, J. A. (2013b). Towards realtime multimodal fusion for image-guided interventions using self-similarities. In *International conference on medical image computing and computer-assisted intervention*, pages 187–194. Springer.

Heinrich, M. P., Papież, B. W., Schnabel, J. A., and Handels, H. (2014). Non-parametric discrete registration with convex optimisation. In *International Workshop on Biomedical Image Registration*, pages 51–61. Springer International Publishing.

Hering, A., Kuckertz, S., Heldmann, S., and Heinrich, M. P. (2019a). Enhancing label-driven deep deformable image registration with local distance metrics for state-of-the-art cardiac motion tracking. In *Bildverarbeitung für die Medizin 2019*, pages 309–314. Springer.

Hering, A., Kuckertz, S., Heldmann, S., and Heinrich, M. P. (2019b). Memory-efficient 2.5 d convolutional transformer networks for multi-modal deformable registration with weak label supervision applied to whole-heart CT and MRI scans. *International journal of computer assisted radiology and surgery*, 14(11):1901–1912.

Hill, D. L., Batchelor, P. G., Holden, M., and Hawkes, D. J. (2001). Medical image registration. *Physics in medicine & biology*, 46(3):R1.

Hlavac, M., Wirtz, C. R., and Halatsch, M.-E. (2017). Intraoperative magnetic resonance imaging. *HNO*, 65(1):25–29.

Hu, S., Kang, H., Baek, Y., El Fakhri, G., Kuang, A., and Choi, H. S. (2018a). Real-time imaging of brain tumor for image-guided surgery. *Advanced healthcare materials*, 7(16):1800066.

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

Hu, Y., Modat, M., Gibson, E., Li, W., Ghavami, N., Bonmati, E., Wang, G., Bandula, S., Moore, C. M., Emberton, M., et al. (2018c). Weakly-supervised convolutional neural networks for multimodal image registration. *Medical image analysis*, 49:1–13.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

Hur, J. and Roth, S. (2016). Joint optical flow and temporally consistent semantic segmentation. In *European Conference on Computer Vision*, pages 163–177. Springer.

Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., and Brox, T. (2017). Flownet 2.0: Evolution of optical flow estimation with deep networks. In *IEEE conference on computer vision and pattern recognition (CVPR)*, volume 2, page 6.

Jaderberg, M., Simonyan, K., Zisserman, A., et al. (2015). Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025.

Keall, P. J., Mageras, G. S., Balter, J. M., Emery, R. S., Forster, K. M., Jiang, S. B., Kapatoes, J. M., Low, D. A., Murphy, M. J., Murray, B. R., et al. (2006). The management of respiratory motion in radiation oncology report of AAPM Task Group 76 a. *Medical physics*, 33(10):3874–3900.

Ker, J., Wang, L., Rao, J., and Lim, T. (2017). Deep learning applications in medical image analysis. *Ieee Access*, 6:9375–9389.

Khan, A., Sohail, A., Zahoora, U., and Qureshi, A. S. (2020). A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review*, 53(8):5455–5516.

King, A., Buerger, C., Tsoumpas, C., Marsden, P., and Schaeffter, T. (2012). Thoracic respiratory motion estimation from MRI using a statistical model and a 2D image navigator. *Medical Image Analysis*, 16(1):252 – 264.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Klein, S., Pluim, J. P., Staring, M., and Viergever, M. A. (2009). Adaptive stochastic gradient descent optimisation for image registration. *International journal of computer vision*, 81(3):227.

Klinder, T. and Lorenz, C. (2012). Respiratory motion compensation for image-guided bronchoscopy using a general motion model. In *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 960–963. IEEE.

Klüter, S. (2019). Technical design and concept of a 0.35 t MR-linac. *Clinical and translational radiation oncology*, 18:98–101.

König, L., Kipshagen, T., and Rühaak, J. (2014). A non-linear image registration scheme for real-time liver ultrasound tracking using normalized gradient fields. In *MICCAI Challenge on Liver Ultrasound Tracking CLUST 2014*, page 29.

Korreman, S. (2015). Image-guided radiotherapy and motion management in lung cancer. *The British journal of radiology*, 88(1051):20150100.

Krebs, J., Mansi, T., Mailhé, B., Ayache, N., and Delingette, H. (2018). Unsupervised probabilistic deformation modeling for robust diffeomorphic registration. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 101–109. Springer.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90.

Landberg, T., Chavaudra, J., Dobbs, J., Hanks, G., Johansson, K.-A., Möller, T., and Purdy, J. (2016). Report 50. *Journal of the International Commission on Radiation Units and Measurements*, os26(1):NP–NP.

Le, V., Brandt, J., Lin, Z., Bourdev, L., and Huang, T. S. (2012). Interactive facial feature localization. In *European conference on computer vision*, pages 679–692. Springer.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Lee, D. and Krupa, A. (2011). Intensity-based visual servoing for non-rigid motion compensation of soft tissue structures due to physiological motion using 4D ultrasound. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2831–2836.

Li, H. and Fan, Y. (2017). Non-rigid image registration using fully convolutional networks with deep self-supervision. *arXiv preprint arXiv:1709.00799*.

Li, H., Kadav, A., Durdanovic, I., Samet, H., and Graf, H. P. (2016). Pruning filters for efficient convnets. *arXiv preprint arXiv:1608.08710*.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B., and Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88.

Liu, Z., Sun, M., Zhou, T., Huang, G., and Darrell, T. (2019). Rethinking the value of network pruning. In *International Conference on Learning Representations*.

Loh, J., Baker, K., Sridharan, S., Greer, P., Wratten, C., Capp, A., Gallagher, S., and Martin, J. (2015). Infections after fiducial marker implantation for prostate radiotherapy: are we underestimating the risks? *Radiation Oncology*, 10(1):1–5.

Lorenzi, M., Ayache, N., Frisoni, G. B., Pennec, X., (ADNI, A. D. N. I., et al. (2013). LCC-Demons: a robust and accurate symmetric diffeomorphic registration algorithm. *NeuroImage*, 81:470–483.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110.

Luca, V. D., Benz, T., Kondo, S., König, L., Lübke, D., Rothlübbers, S., Somphone, O., Allaire, S., Bell, M. A. L., Chung, D. Y. F., Cifor, A., Grozea, C., Günther, M., Jenne, J., Kipshagen, T., Kowarschik, M., Navab, N., Rühaak, J., Schwaab, J., and Tanner, C. (2015). The 2014 liver ultrasound tracking benchmark. *Physics in Medicine and Biology*, 60(14):5571.

Luo, X. and Zhuang, X. (2020). Mvmm-regnet: A new image registration framework based on multivariate mixture model and neural network estimation. *arXiv preprint arXiv:2006.15573*.

Machado, I., Toews, M., Luo, J., Unadkat, P., Essayed, W., George, E., Teodoro, P., Carvalho, H., Martins, J., Golland, P., Pieper, S., Frisken, S., Golby, A., Wells III, W., and Ou, Y. (2018). Deformable MRI-Ultrasound Registration via Attribute Matching and Mutual-Saliency Weighting for Image-Guided Neurosurgery. In Stoyanov, D., Taylor, Z., Aylward, S., Tavares, J. M. R., Xiao, Y., Simpson, A., Martel, A., Maier-Hein, L., Li, S., Rivaz, H., Reinertsen, I., Chabanas, M., and Farahani, K., editors, *Simulation, Image Processing, and Ultrasound Systems for Assisted Diagnosis and Navigation*, pages 165–171, Cham. Springer International Publishing.

Maintz, J. A. and Viergever, M. A. (1998). A survey of medical image registration. *Medical image analysis*, 2(1):1–36.

Maurer, C. R. and Fitzpatrick, J. M. (1993). A review of medical image registration. *Interactive image-guided neurosurgery*, 17.

McClelland, J. R., Hawkes, D. J., Schaeffter, T., and King, A. P. (2013). Respiratory motion models: a review. *Medical image analysis*, 17(1):19–42.

Mittauer, K., Paliwal, B., Hill, P., Bayouth, J. E., Geurts, M. W., Baschnagel, A. M., Bradley, K. A., Harari, P. M., Rosenberg, S., Brower, J. V., et al. (2018). A new era of image guidance with magnetic resonance-guided radiation therapy for abdominal and thoracic malignancies. *Cureus*, 10(4).

Modersitzki, J. (2004). *Numerical methods for image registration*. Oxford University Press on Demand.

Mutic, S. (2012). WE-A-BRA-02: First commercial hybrid MRI-IMRT system. *Medical Physics*, 39(6Part25):3934–3934.

Bibliography

Myronenko, A. and Song, X. (2010). Point set registration: Coherent point drift. *IEEE transactions on pattern analysis and machine intelligence*, 32(12):2262–2275.

Otazo, R., Lambin, P., Pignol, J.-P., Ladd, M. E., Schlemmer, H.-P., Baumann, M., and Hricak, H. (2020). MRI-guided radiation therapy: An emerging paradigm in adaptive radiation oncology. *Radiology*, page 202747.

Paganelli, C., Seregni, M., Fattori, G., Summers, P., Bellomi, M., Baroni, G., and Riboldi, M. (2015). Magnetic resonance imaging–guided versus surrogate-based motion tracking in liver radiation therapy: A prospective comparative study. *International Journal of Radiation Oncology*Biology*Physics*, 91(4):840 – 848.

Preiswerk, F., De Luca, V., Arnold, P., Celicanin, Z., Petrusca, L., Tanner, C., Bieri, O., Salomir, R., and Cattin, P. C. (2014). Model-guided respiratory organ motion prediction of the liver from 2D ultrasound. *Medical image analysis*, 18(5):740–751.

Prim, R. C. (1957). Shortest connection networks and some generalizations. *The Bell System Technical Journal*, 36(6):1389–1401.

Qin, C., Bai, W., Schlemper, J., Petersen, S. E., Piechnik, S. K., Neubauer, S., and Rueckert, D. (2018). Joint learning of motion estimation and segmentation for cardiac MR image sequences. *arXiv preprint arXiv:1806.04066 (MICCAI 2018)*.

Raghu, M., Zhang, C., Kleinberg, J., and Bengio, S. (2019). Transfusion: Understanding transfer learning for medical imaging. In *Advances in neural information processing systems*, pages 3347–3357.

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

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

Royer, L., Krupa, A., Dardenne, G., Bras, A. L., Marchand, E., and Marchal, M. (2017). Real-time target tracking of soft tissues in 3D ultrasound images based on robust visual information and mechanical simulation. *Medical Image Analysis*, 35:582 – 598.

Ruan, D., Fessler, J. A., Balter, J., and Keall, P. (2009). Real-time profiling of respiratory motion: baseline drift, frequency variation and fundamental pattern change. *Physics in Medicine & Biology*, 54(15):4777.

Rueckert, D., Aljabar, P., Heckemann, R. A., Hajnal, J. V., and Hammers, A. (2006). Diffeomorphic registration using b-splines. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 702–709. Springer.

Rueckert, D., Sonoda, L. I., Hayes, C., Hill, D. L., Leach, M. O., and Hawkes, D. J. (1999). Nonrigid registration using free-form deformations: application to breast MR images. *IEEE transactions on medical imaging*, 18(8):712–721.

Ruhaak, J., Polzin, T., Heldmann, S., Simpson, I., Handels, H., Modersitzki, J., and Heinrich, M. P. (2017). Estimation of large motion in lung CT by integrating regularized keypoint correspondences into dense deformable registration. *IEEE Transactions on Medical Imaging*, 36:1746–1757.

Saenz, D. L., Astorga, N. R., Kirby, N., Fakhreddine, M., Rasmussen, K., Stathakis, S., and Papanikolaou, N. (2018). A method to predict patient-specific table coordinates for quality assurance in external beam radiation therapy. *Journal of applied clinical medical physics*, 19(5):625–631.

Sastry, R., Bi, W. L., Pieper, S., Frisken, S., Kapur, T., Wells III, W., and Golby, A. J. (2017). Applications of ultrasound in the resection of brain tumors. *Journal of Neuroimaging*, 27(1):5–15.

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

Seregni, M., Paganelli, C., Summers, P., Bellomi, M., Baroni, G., and Riboldi, M. (2017, in press). A hybrid image registration and matching framework for real-time motion tracking in MRI-guided radiotherapy. *IEEE Transactions on Biomedical Engineering*.

Sevilla-Lara, L., Sun, D., Jampani, V., and Black, M. J. (2016). Optical flow with semantic segmentation and localized layers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3889–3898.

Shams, R., Boucher, M.-A., and Kadoury, S. (2018). Intra-operative brain shift correction with weighted locally linear correlations of 3DUS and MRI. In *Simulation, Image Processing, and Ultrasound Systems for Assisted Diagnosis and Navigation*, pages 179–184. Springer.

Shepard, A. J., Wang, B., Foo, T. K., and Bednarz, B. P. (2017, in press). A block matching based approach with multiple simultaneous templates for the real-time 2D ultrasound tracking of liver vessels. *Medical Physics*, pages n/a–n/a.

Sirignano, J. and Spiliopoulos, K. (2018). DGM: A deep learning algorithm for solving partial differential equations. *Journal of computational physics*, 375:1339–1364.

Smith, B. M., Zhang, L., Brandt, J., Lin, Z., and Yang, J. (2013). Exemplar-based face parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3484–3491.

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

Somphone, O., Allaire, S., Mory, B., and Dufour, C. (2014). Live feature tracking in ultrasound liver sequences with sparse demons. In *MICCAI Challenge on Liver Ultrasound Tracking CLUST 2014*, pages 53–60.

Sotiras, A., Davatzikos, C., and Paragios, N. (2013). Deformable medical image registration: A survey. *IEEE transactions on medical imaging*, 32(7):1153–1190.

Srivastava, R. K., Greff, K., and Schmidhuber, J. (2015). Training very deep networks. In *Advances in neural information processing systems*, pages 2377–2385.

Bibliography

Stemkens, B., Tijssen, R., de Senneville, B., Lagendijk, J., and van den Berg, C. (2016). Image-driven, model-based 3D abdominal motion estimation for MR-guided radiotherapy. *Physics in Medicine and Biology*, 61(14):5335.

Tanner, C., Ozdemir, F., Profanter, R., Vishnevsky, V., Konukoglu, E., and Goksel, O. (2018). Generative adversarial networks for MR-CT deformable image registration. *arXiv preprint arXiv:1807.07349*.

Tsai, Y.-H., Yang, M.-H., and Black, M. J. (2016). Video segmentation via object flow. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3899–3908.

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

Wang, W., Dang, Z., Hu, Y., Fua, P., and Salzmann, M. (2019). Backpropagation-friendly eigendecomposition. In *Advances in Neural Information Processing Systems*, pages 3162–3170.

Wein, W. (2018). Brain-shift correction with image-based registration and landmark accuracy evaluation. In *Simulation, Image Processing, and Ultrasound Systems for Assisted Diagnosis and Navigation*, pages 146–151. Springer.

Wilms, M. (2018). *Effiziente Schätzung der Atembewegung mittels statistischer Bewegungs- und regressionbasierter Korrespondenzmodell*. dissertation, Universität zu Luebeck.

Wilms, M., Ha, I. Y., Handels, H., and Heinrich, M. P. (2016). Model-based regularisation for respiratory motion estimation with sparse features in image-guided interventions. In Ourselin, S., Joskowicz, L., Sabuncu, M. R., Unal, G., and Wells, W., editors, *19th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) 2016*, volume 9902 of *Image Processing, Computer Vision, Pattern Recognition, and Graphics*, pages 89–97, Athen. Springer International Publishing, Springer International Publishing.

Winkel, D., Bol, G. H., Kroon, P. S., van Asselen, B., Hackett, S. S., Werensteijn-Honingh, A. M., Intven, M. P., Eppinga, W. S., Tijssen, R. H., Kerkmeijer, L. G., et al. (2019). Adaptive radiotherapy: the Elekta Unity MR-linac concept. *Clinical and translational radiation oncology*, 18:54–59.

Xiao, Y., Rivaz, H., Chabanas, M., Fortin, M., Machado, I., Ou, Y., Heinrich, M. P., Schnabel, J. A., Zhong, X., Maier, A., et al. (2019). Evaluation of MRI to ultrasound registration methods for brain shift correction: The CuRIOUS2018 Challenge. *IEEE Transactions on Medical Imaging*.

Xu, Z., Lee, C. P., Heinrich, M. P., Modat, M., Rueckert, D., Ourselin, S., Abramson, R. G., and Landman, B. A. (2016). Evaluation of six registration methods for the human abdomen on clinically acquired ct. *IEEE Transactions on Biomedical Engineering*, 63(8):1563–1572.

Xu, Z., Luo, J., Yan, J., Pulya, R., Li, X., Wells III, W., and Jagadeesan, J. (2020). Adversarial uni-and multi-modal stream networks for multimodal image registration. *arXiv preprint arXiv:2007.02790*.

Yang, Q., Wang, L., and Ahuja, N. (2010). A constant-space belief propagation algorithm for stereo matching. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1458–1465. IEEE.

Yoo, I., Hildebrand, D. G., Tobin, W. F., Lee, W.-C. A., and Jeong, W.-K. (2017). ssemnet: Serial-section electron microscopy image registration using a spatial transformer network with learned features. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 249–257. Springer.

Zachiu, C., Papadakis, N., Ries, M., Moonen, C., and De Senneville, B. D. (2015). An improved optical flow tracking technique for real-time MR-guided beam therapies in moving organs. *Physics in Medicine & Biology*, 60(23):9003.

Zhong, X., Bayer, S., Ravikumar, N., Strobel, N., Birkhold, A., Kowarschik, M., Fahrig, R., and Maier, A. (2018). Resolve intraoperative brain shift as imitation game. In *Simulation, Image Processing, and Ultrasound Systems for Assisted Diagnosis and Navigation*, pages 129–137. Springer.