Lehrstuhl für Computerunterstützte Klinische Medizin
der Medizinischen Fakultät Mannheim, Universität Heidelberg
(Direktor: Prof. Dr. rer. nat. Lothar R. Schad)

Methods for three-dimensional Registration of Multimodal Abdominal Image Data

Inauguraldissertation
zur Erlangung des
Doctor scientiarum humanarum (Dr. sc. hum.)
der
Medizinischen Fakultät Mannheim
der Ruprecht-Karls-Universität
zu
Heidelberg

vorgelegt von
Barbara Ingeborg Waldkirch

aus
Ludwigshafen am Rhein

2020

Dean: Prof. Dr. med. Sergij Goerdt
Supervisor: Prof. Dr. rer. nat. Lothar R. Schad

Abstract

Multimodal image registration benefits the diagnosis, treatment planning and the performance of image-guided procedures in the liver, since it enables the fusion of complementary information provided by pre- and intrainterventional data about tumor localization and access. Although there exist various registration methods, approaches which are specifically optimized for the registration of multimodal abdominal scans are only scarcely available. The work presented in this thesis aims to tackle this problem by focusing on the development, optimization and evaluation of registration methods specifically for the registration of multimodal liver scans.

The contributions to the research field of medical image registration include the development of a registration evaluation methodology that enables the comparison and optimization of linear and non-linear registration algorithms using a point-based accuracy measure. This methodology has been used to benchmark standard registration methods as well as novel approaches that were developed within the frame of this thesis. The results of the methodology showed that the employed similarity measure used during the registration has a major impact on the registration accuracy of the method.

Due to this influence, two alternative similarity metrics bearing the potential to be used on multimodal image data are proposed and evaluated. The first metric relies on the use of gradient information in form of *Histograms of Oriented Gradients* (HOG) whereas the second metric employs a siamese neural network to learn a similarity measure directly on the image data. The evaluation showed, that both metrics could compete with state of the art similarity measures in terms of registration accuracy. The HOG-metric offers the advantage that it does not require ground truth data to learn a similarity estimation, but instead it is applicable to various data sets with the sole requirement of distinct gradients. However, the Siamese metric is characterized by a higher robustness for large rotations than the HOG-metric. To train such a network, registered ground truth data is required which may be critical for multimodal image data. Yet, the results show that it is possible to apply models trained on registered synthetic data on real patient data.

The last part of this thesis focuses on methods to learn an entire registration process using neural networks, thereby offering the advantage to replace the traditional, time-consuming iterative registration procedure. Within the frame of this thesis, the so-called *VoxelMorph* network which was originally proposed for monomodal, non-linear registration learning is extended for affine and multimodal registration learning tasks. This extension includes the consideration of an image mask during metric evaluation as well as loss functions for multimodal data, such as the pretrained Siamese metric and a loss relying on the comparison of deformation fields. Based on the developed registration evaluation methodology, the performance of the original network as well as the extended variants are evaluated for monomodal and multimodal registration tasks using multiple data sets. With the extended network variants, it is possible to learn an entire multimodal registration process for the correction of large image displacements. As for the Siamese metric, the results imply a general transferability of models trained with synthetic data to registration tasks including real patient data. Due to the lack of multimodal ground truth data, this transfer represents an important step towards making Deep Learning based registration procedures clinically usable.

Zusammenfassung

Multimodale Bildregistrierung ist ein wichtiges Forschungsgebiet in der medizinischen Bildverarbeitung, da sie die Fusion von komplementären Informationen ermöglicht, die von verschiedenen bildgebenden Verfahren geliefert werden. Obwohl es eine Vielzahl an Registrierungsmethoden gibt, sind Ansätze, die speziell für die Registrierung von multimodalen Bildaufnahmen des Abdomens optimiert wurden, kaum verfügbar. Das Ziel dieser Arbeit ist deshalb die Entwicklung, Optimierung und Evaluation von Verfahren für die Registrierung multimodaler Bilddaten des Abdomens.

Die vorgeschlagenen Beiträge zum Forschungsgebiet der medizinischen Bildregistrierung umfassen die Entwicklung einer Evaluationsmethodik für Registrierungsverfahren, die den Vergleich und die Optimierung von linearen und nichtlinearen Registrierungsmethoden mittels eines punktbasierten Genauigkeitsmaßes ermöglicht. Diese Methodik wurde sowohl für die Bewertung und Optimierung von Standardregistrierungsmethoden als auch für neuartige Ansätze, die in dieser Arbeit entwickelt wurden, verwendet. Die Ergebnisse zeigen, dass vor allem die für die Registrierung verwendete Ähnlichkeitsmetrik einen großen Einfluss auf die Registrierungsgenauigkeit der Methode hat.

Daher wurden im Rahmen dieser Arbeit zwei alternative Ähnlichkeitsmetriken für den Vergleich von multimodale Bilddaten entwickelt und evaluiert. Die erste Metrik beruht auf der Verwendung von Gradienteninformation in Form von *Histogrammen Orientierter Gradienten* (HOG), während die zweite Metrik ein sog. Siamesisches neuronales Netz verwendet, um ein Ähnlichkeitsmaß direkt auf den vorliegenden Bilddaten zu erlernen. Die Auswertung zeigt, dass beide Metriken in Bezug auf die erreichbare Registrierungsgenauigkeit mit traditionellen Ähnlichkeitsmaßen konkurrieren können. Die HOG-Metrik bietet den Vorteil, dass sie zum Erlernen einer Ähnlichkeitsschätzung keine Ground Truth-Daten benötigt, sondern auf verschiedensten Datensätzen anwendbar ist. Allerdings zeichnet sich die siamesische Metrik durch eine höhere Robustheit für große Rotationen aus. Nachteil eines siamesischen Netzwerks ist der Bedarf an registrierten Ground Truth-Daten um das Netz trainieren zu können. Die Ergebnisse weisen jedoch auf eine allgemeine Anwendbarkeit der mit synthetischen Daten trainierten Modelle auf realen Patientendaten hin.

Der letzte Teil dieser Arbeit konzentriert sich auf die Verwendung von neuronalen Netzen, um einen kompletten Registrierungsprozess zu erlernen. In dieser Arbeit wurde das sog. *VoxelMorph*-Netzwerk, das ursprünglich für das Erlernen eines monomodalen, nichtlinearen Registrierungsprozesses vorgestellt wurde, für affine und multimodale Registrierungsaufgaben erweitert. Diese Erweiterung beinhaltet die Berücksichtigung einer Bildmaske bei der Metrikberechnung, sowie die Integration alternativer Verlustfunktionen die auf multimodalen Daten anwendbar sind. Diese Funktionen umfassen die vortrainierte Siamesische Metrik, sowie eine Verlustfunktion, die auf dem Vergleich von Deformationsfeldern beruht. Basierend auf der entwickelten Evaluationsmethodik wurde die Registrierungsgenauigkeit des ursprünglichen Netzes sowie der erweiterten Varianten für monomodale und multimodale Registrierungen bewertet. Die Ergebnisse zeigen, dass es mit den erweiterten Netzvarianten möglich ist, einen Registrierungsprozess für die Korrektur großer Bildverschiebungen zu erlernen. Des Weiteren zeigen die Resultate auch hier eine Übertragbarkeit der mit synthetischen Daten trainierten Modelle auf die Registrierung realer Patientendaten. Aufgrund des Mangels multimodaler Ground Truth-Daten, repräsentiert dieser Transfer einen ersten Schritt um auf Deep Learning basierende Registrierungsverfahren klinisch nutzbar zu machen.

List of Abbreviations

AMMI	Advanced Mattes Mutual Information
CBCT	Cone-Beam Computed Tomography
CNN	Convolutional Neural Network
CT	Computed Tomography
DL	Deep Learning
DFL	Deformation Field Loss
DOF	Degrees Of Freedom
DSC	Dice Similarity Coefficient
FCN	Fully Convolutional Network
FCSS	Fully Convolutional Self-Similarity
FLE	Fiducial Localization Error
FOV	Field of View
FRE	Fiducial Registration Error
GAN	Generative Adversarial Network
GPU	Graphics Processing Unit
GT	Ground Truth
HOG	Histogram of Oriented Gradients
ITK	Insight Segmentation and Registration Toolkit
M^2OLIE	Mannheim Molecular Intervention Environment
MI	Mutual Information
MIND	Modality Independent Neighborhood Descriptor
MITK	Medical Imaging Interaction Toolkit
MRI	Magnetic Resonance Imaging
MSE	Mean-Squared Error
NCC	Normalized Cross-Correlation

NGF	Normalized Gradient Field
NMI	Normalized Mutual Information
OMD	Oligometastatic Disease
PET	Positron Emission Tomography
ROI	Region Of Interest
RF	Radio Frequency
SDM	Siamese Deep Metric
SIFT	Scale-Invariant Feature Transform
SSD	Sum of Squared Differences
STN	Spatial Transformer Network
SURF	Speeded Up Robust Features
TACE	Transarterial Chemoembolization
TRE	Target Registration Error
2D	Two-dimensional
3D	Three-dimensional

Contents

1

Introduction

1.1 Motivation

Multimodal image registration is an important research field in medical image processing, since it enables the fusion of complementary information provided by different imaging modalities. This fusion can benefit various applications in the clinical context, ranging from an improved diagnosis, to treatment planning and navigation during image-guided procedures. Image registration can even be relevant for treatment monitoring, since it e.g. enables the control of tumor shrinkage after treatment, by registration of pre- and post-procedural image data.

Preoperative imaging data is used for the visualization and localization of tumor tissue and anatomical landmarks and therefore builds the basis for interventional and surgical planning. It is often useful to acquire image data of more than one modality, since different image modalities provide different information. However, the precise localization of target anatomy with respect to planning data during an intervention or surgery can be challenging due to differences in patient positioning, anatomical deformations and the intervention itself. Therefore, intra-interventional image data is additionally acquired in order to visualize contrasted vessels, interventional tools like a catheter or the patients anatomy for navigation. Image registration then aims to incorporate the intraoperative information as a real-time update in the preoperative surgical planning. Thus, image registration is not only an important means for preoperative interventional/surgical planning, but can also benefit other applications such as automatized tool positioning and tracking during the intervention.

Exemplary use-cases for such a scenario are image-guided procedures in the liver, such as e.g. biopsies or treatment of liver cancer by means of a transarterial chemoembolization. The improvement of the whole treatment cycle of oligometastatic liver cancer, including the performance of image-guided procedures in the liver, is subject of the research campus "Mannheim Molecular Intervention Environment" (M^2OLIE). Following the intended workflow of M^2OLIE, the pre- and post-interventional data is commonly acquired using three-dimensional computed tomography (CT) and magnetic resonance imaging (MRI), whereas the intraoperative data corresponds to projective X-ray fluoroscopy or three-dimensional cone-beam computed tomography (CBCT). The aim is to fuse all available morphological, functional and molecular information about the tumor provided by the different imaging modalities applied prior to the intervention into a multimodal data set. This data set then builds the basis for an individualized treatment and is ultimately transferred in

the intervention room and registered to the interventional image data to improve surgical planning and the guidance of a robotic assistance system. Hence, such a scenario is based on the use of multimodal image registration methods. There are various requirements for registration methods in this context ranging from the use of an appropriate similarity metric which is able to compare the image data of different modalities, to fast computation times to provide a time-efficient registration of pre- and intraoperational data. Moreover, registration methods for interventional procedures are required to yield a high registration accuracy that results in an optimal overlap of corresponding structures in the images.

In general, the quality of the registration can have a high impact on the interventional or surgical outcome, since a clear definition of the tumor margins is essential for its localization during an intervention. Especially for soft tissues, this task can be very challenging due to tissue deformation. The registration of the liver represents a particular difficult task, since the liver tissue is deformed by respiratory as well as digestive motion. Yet, there exists no universal solution to image registration and registration methods have to be optimized for a specific task. Due to the challenges of registering the liver, methods, especially multimodal registration methods, which are specifically designed for this task are only scarcely available [1, 2].

Therefore, the work presented in this thesis focuses on the development, evaluation and optimization of image registration approaches for three-dimensional multimodal scans of the liver with regard to interventional applications.

1.2 Objectives

As a result of the limited availability of specified registration methods for abdominal image data, general-purpose registration methods, which were often designed and optimized for other body parts (mostly the brain), are transferred to be used for the registration of abdominal scans [3]. This potentially leads to suboptimal results for a number of reasons, including i.e. different degrees of organ displacement, different image resolutions depending on the utilized imaging techniques, but foremost different intensity distributions in the images which are to be registered. Additionally, the task of registering the liver entails specific challenges, due to the homogeneity of its tissue displaying less structure than i.e. the brain. Therefore, the choice and parametrization of the registration method for abdominal scans reveals itself to be a challenging task and a performance characterization of different methods to attain an accurate and robust registration result can be necessary.

The general aim of this thesis is the development of novel multimodal registration methods for abdominal data, as well as the optimization of existing methods for this task. The focus is set on use-cases related to image-guided procedures of the liver, and therefore, the methods presented in this work focus on the registration of three-dimensional (3D) CT, MRI and CBCT data.

Yet, to optimize a registration method and characterize its performance, an evaluation standard has to be defined. This is a critical point in the field of image registration, since there exists no standardized evaluation methodology to compare and benchmark registration methods until today.

This is mainly caused by the diversity of different registration types [4] and the lack of appropriate ground truth data that corresponds to optimally registered data for inter-subject or multimodal registration.

Therefore, the first part of this thesis aims to develop an evaluation methodology that enables a comparison and optimization of linear and non-linear multimodal registration methods based on the estimation of registration accuracy. In general, the development of an evaluation strategy requires a set of registered ground truth data as reference for an optimal image alignment. However, these data sets are only scarcely available, especially for multimodal data, due to differences in the acquisition procedures of different imaging modalities. Therefore, the development of an evaluation methodology also includes the generation of multimodal ground truth data sets. Thus, different approaches for the generation of these data sets are investigated, including manual preprocessing of actual patient data as well as the generation of synthetic ground truth data using a neural network. The idea is to implement an evaluation methodology in form of a framework that allows the integration of any available registration approach. This methodology can then be used to evaluate and optimize the performance of state of the art registration methods as well as novel approaches for the multimodal registration of abdominal scans.

The most challenging part in multimodal image registration is represented by the choice of an appropriate similarity measure to estimate the alignment of corresponding structures in images of different modality. The main difficulty is represented by the fact, that different image acquisition techniques may result in very dissimilar grey value distributions in the images, so that the same anatomical features appear differently in these images. This potentially leads to the generation of statistical correlations between image structures that do not correspond to the same anatomical structures, thus violating the main assumption of most intensity-based similarity measures. Up to now, there exists only a limited number of multimodal similarity measures that mostly rely on the concept of mutual information [5]. To investigate further alternatives, the second part of the thesis focuses on the development and evaluation of similarity measures that bear the potential to be applicable on multimodal image data.

The last part of the thesis aims to investigate approaches to perform end-to-end registration learning using a neural network. The rise of Deep Learning methods in the field of medical image processing benefits image registration not only in terms of novel similarity estimations but also in the development of neural networks that are able to learn an entire registration process. Since traditional registration methods correspond to iterative optimization procedures, an advantage of novel approaches using Deep Learning is the fast computation time of the registration process, once such a network is trained. A widely used network architecture for end-to-end registration learning is the so-called VoxelMorph network [6]. However, as most of these registration learning networks, the VoxelMorph network is currently restricted to monomodal nonlinear registration learning. Therefore, the last part of this thesis aims at the extension of the VoxelMorph network for the application on multimodal image data as well as for affine registration to enable the correction of larger image displacements.

In summary, this thesis aims to contribute to three different research areas in the field of medical image registration:

- Image registration evaluation and generation of ground truth data,
- novel similarity measures,
- and Deep Learning in medical image registration.

1.3 Thesis Structure

This thesis is composed of seven parts. After this introductory **chapter 1** explaining the motivation for the work presented in this thesis, all theoretical basics which are relevant to this work will be explained in **chapter 2**. This includes a short presentation of the medical background of image-guided procedures in the liver. Since this work focuses on the image registration of multimodal data, chapter 2 also includes an explanation of the physical principals for the image contrast generation in CT, CBCT and MRI. Next, the fundamentals of medical image registration such as the different types of registration, the main components of an registration algorithm as well as the challenges and limitations for registration methods are explained. Since this work not only relies on traditional image processing but also on novel approaches using artificial neural networks, chapter 2 also includes a general overview of the fundamentals of Deep Learning. This work proposes contributions to three different research areas in the field of medical image registration: image registration evaluation, novel feature-based similarity metrics and Deep Learning in medical image registration. The current state of the art of these three topics and a brief description of the advancements proposed in this thesis are presented in **chapter 3**. In **chapter 4**, the approches proposed in this work are presented in detail. This includes the presentation of the ground truth data used for the experiments in this thesis and the developed evaluation methodology for linear and non-linear registration methods. Moreover, the basics and implementation details of two alternative similarity metrics relying on traditional HOG features and a siamese network are presented as well as the extensions integrated in the VoxelMorph network to enable affine and multimodal end-to-end registration learning. The results obtained for the evaluation of various registration methods, including registrations based on the alternative similarity measures as well as the extended VoxelMorph network, using the novel evaluation methodology, are presented in **chapter 5** and discussed in detail in **chapter 6**. The thesis concludes with a summary of the work presented in this thesis and an outlook to future work concerning the proposed approaches given in **chapter 7**.

2

Theoretical Background

In the following chapter, the medical background and the clinical use-cases of this work are described. Moreover, the fundamentals of medical image acquisition as well as medical image registration are presented. The content and characteristics of medical images highly depend on the acquisition technique. Therefore, the following section is dedicated to create a deeper understanding of the fundamentals of medical image acquisition causing the different contrasts in different modalities.

Since the focus of this thesis is set on the development and evaluation of image registration methods, an overview over the general structure of image registration algorithms is presented. The basics for each algorithm are explained as well as the general challenges of image registration. To take into account the latest developments in the field of medical image processing, the fundamentals of deep neural networks and their influence on the field of medical image registration will also be presented.

2.1 Medical Background

This work aims to develop and optimize image registration algorithms specifically for the registration of abdominal scans that are acquired for the diagnosis and treatment of liver cancer. To further understand the necessity of image registration in this context, the following section aims to described the anatomy of the human liver as well as common diseases and the treatment focusing on image-guided interventional procedures.

Anatomy of the Human Liver

The liver is the largest internal organ in the human body and occupies a multitude of important and complex functions concerning the entire human metabolism. It is located in the right upper part of the abdomen, partially covered by the ribcage right below the diaphragm. The human liver typically weights between 1.5 and 2 kg, making it the heaviest organ, and the largest gland in the body of vertebrates [7]. The anatomy of the liver is divided into two main sections, the right lobe (*Lobus Dexter*) and left lobe (*Lobus sinister*) in axial view, which are again subdivided in 8 subsegments according to the *Couinaud system*. This system relies relies on functional anatomy and divides the liver in 8 parts based on a transverse plane through the bifurcation of the portal vein

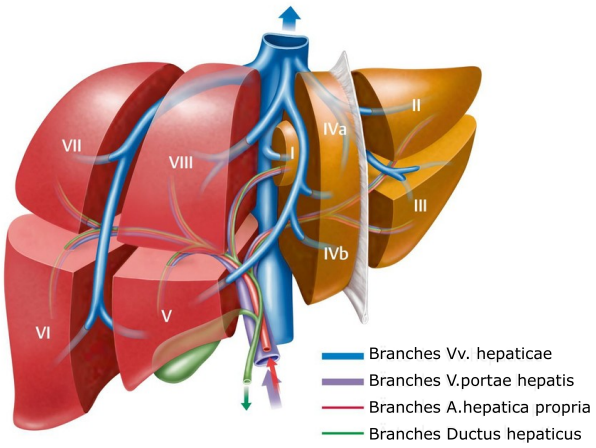


Figure 2.1: Schematic illustration of the human liver, displaying the vascular system and the liver segments according to the Couinaud system (image adapted from [9]).

(*Porta hepatis*) as shown in fig. 2.1. In the hepatic portal system, the liver receives a double blood supply: the portal vein carries venous blood from the gastrointestinal tract to the liver providing 70% of the total liver blood flow, while 30% comes from the right and left hepatic arteries which carry oxygenated blood to the liver [8]. The liver regulates a variety of different vital functions including detoxification, synthesis and storage. It i.a. filters and removes harmful substances and toxins from the body, assures the metabolism of carbohydrates, fats and proteins while producing bile which is essential for the digestion process and stores glucose, vitamins and iron.

Liver Diseases

In general, the liver is prone to many diseases. These range from diseases caused by viruses, such as hepatitis, to diseases caused by intoxication such as the fatty liver disease and cirrhosis, inherited diseases such as hemochromatosis or Wilson disease and liver cancer. In some cases, the development of liver cancer can be linked to cirrhosis, which describes an increase in the scarring of the liver tissue caused by a previous hepatitis infection or hemochromatosis [10], but the general causing effects for the development of a liver tumor are not known. Since this work mainly focuses on use-cases for the diagnosis and treatment of liver cancer, the different types of liver cancer are described in more detail in the following.

It is distinguished between primary and secondary liver cancer. Primary tumors grow at the organ where the tumor progression began whereas secondary tumors, so-called metastasis, are caused by the spread of cancer cells from a primary site to a secondary site within surrounding tissue or to a distant organ by intruding the circulatory or lymphathic system. Metastases are the major cause of cancer morbidity and mortality and it is estimated that they account for 90% of cancer deaths [11]. The most frequent sites for the spread of metastasis are lungs, liver, brain and bones [12]. The most common type of primary liver cancer in adults is the hepatocellular carcinoma (HCC), a malignant liver tumor which ranks as the second leading cause of cancer death in East Asia and sub-Saharan Africa and the sixth most common cause of cancer death in western countries [13].

Due the blood supply from the abdominal organs into the liver via portal vein, tumor cells can

spread from these organs into the liver parenchyma, thus making the liver also prone to be a site for metastatic (secondary) cancer. Liver metastases are often linked to colorectal cancer metastasis which is the second most common tumor type in Germany [14]. If the tumor metastasizes only to a limited number of sites and number of distinct metastases (typically between 1 and 5 metastases), the disease is referred to as oligometastatic disease (OMD). Although metastatic malignancies are generally associated with a poor treatment prognosis, the curability highly depends on the number and diameter of the metastases, thus increasing the possibility to successfully treat OMD [15].

Image-guided Procedures in the Liver

To determine the severity of the liver disease, the extraction and analysis of samples cells or tissues by means of a biopsy is an important measure for diagnosis and treatment monitoring [16]. A percutaneous liver biopsy involves the insertion of a thin biopsy needle through the patients abdomen to extract a small tissue sample which can then be analyzed on a molecular level. Other types of biopsy include the transjugular biopsy, during which the needle is inserted via catheter and a vein in the neck, or the laparoscopic biospy using a small abdominal incision to enter the needle and endoscope. The analysis of the biopsy then builds the basis for further treatment decisions.

The main focus of treatment for liver metastases is systemic therapy, however local therapies provide an additional possibility to extend survival. These local treatments include surgical resection of liver tumors, which can be challenging due to the location or extend of disease. Additionally, they include liver directed therapies such as transarterial chemoembolization, radio embolization, radiofrequency ablation, microwave ablation and stereotactic body radiotherapy [17]. A short discussion of treatment procedures with respect to the staging of liver cancer is given in [18].

A uniting factor for the diagnostic biopsy as well as the liver directed therapies is that they all represent interventional procedures which are performed using image-guidance in the operation room. In general, image-guided interventions rely on computer-aided systems for the visualization of target and risk structures in the intervention room. The first step for this visualization is often represented by the preprocessing of the diagnostic image data that is acquired before the intervention to develop specific models that can be employed for patient-individual treatment planning as well as guidance during the intervention [19]. These models range from geometrical models enhancing the morphological information (e.g. deformation fields or segmentation masks), functional models (e.g. perfusion maps) to interaction maps of the tissue with radiation or drugs [20].

The basis for most of these applications is the generation of a multimodal image map in which the same organ structures overlap and which displays the complementary information offered by different imaging modalities. The key technology for the generation of a fused data set is *image registration*. Image registration generally describes the determination of a geometric transformation field which transfers two datasets in the same coordinate system in a way that same structures overlay in both datasets. To further understand the necessity of multimodal imaging, the fundamental basics and differences in information content of standard imaging techniques will be further explained in the following section.

2.2 Medical Image Acquisition

Medical imaging refers to techniques and processes which enable a visualization of the interior of a body for diagnosis, treatment and monitoring of medical conditions. There exist various types, also called modalities, of medical imaging procedures, which use different technologies and therefore yield different specific information about the imaged body region. These can range from morphological to functional and sometimes even molecular information. Thus, it is often necessary to combine different imaging techniques to obtain the full understanding of a medical condition. The presented image acquisition techniques are restricted to the modalities which are in standard use for the diagnosis of liver diseases and the performance of liver interventions relying on a C-arm system to acquire interventional image data. The presented modalities comprise Computed Tomography (CT), Cone-Beam Computed Tomography (CBCT) and Magnetic Resonance Imaging (MRI).

2.2.1 X-ray Computed Tomography

In 1895, Wilhem Conrad Roentgen published a radiography of his wife's hand representing the first published medical image [21]. The generation of this image relies on X-ray radiation, a form of high-energy electromagnetic radiation whose wavelength ranges between 0.01 to 10 nm and is therefore not included in the visible spectrum of electromagnetic waves.

There are two different physical processes causing the generation of X-rays: During the first process, a charged particle is decelerated due to the strong electric field near other charged particles, typically an electron which is decelerated by an atomic nuclei. The loss of kinetic energy of the moving particle is converted to electromagnetic radiation displaying a continuous energy spectrum, so called *Bremsstrahlung*. The second process occurs when an orbital electron is knocked out of an atom by a charged particle and the resulting vacancy is filled by an outer-shell electron. The change in energy during this electron transition is compensated by the emission of a quantized photon whose energy corresponds to the energy difference between the higher and the lower orbital level. Due to the discrete energy levels in an atom, this process results in a discrete energy spectrum which depends on the material of the target that is hit by the charged particle. Thus, this type of radiation is called *Characteristic X-rays*.

When X-rays penetrate matter, both the amplitude and the phase of the electromagnetic wave are affected due to different interaction processes taking place in the material. By detecting and analyzing the intensity profile of the outcoming X-rays, it is possible to draw conclusions about the object properties. This builds the basis of X-ray imaging.

In a homogenous object or body, the initial intensity I_0 of monochromatic X-rays is decreased following the so-called Lambert-Beers-Law

$$I(x) = I_0 \cdot e^{-\mu(E,Z)\cdot x} \ , \tag{2.1}$$

whereas x represents the propagation distance in the body and $\mu(E, Z)$ corresponds to the so-

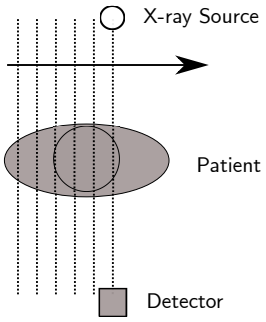


Figure 2.2: Setup for the acquisition of a radiography.

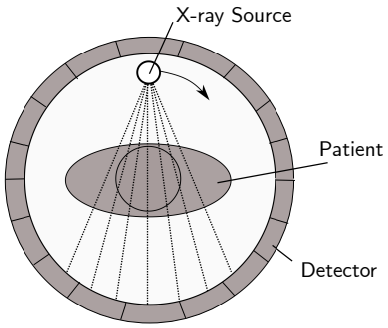


Figure 2.3: Modern Spiral-CT scanners are equipped with detector arrays surrounding the patient 360°.

called linear attenuation coefficient which depends on the X-ray energy E and the atomic number Z of the element that composes the material. The linear attenuation coefficient characterizes the penetrability of a medium.

The fact that X-ray radiation is attenuated in a body can be used for imaging purposes by aligning an X-ray source with a patient and a detector to capture the resulting intensity profile as shown in Figure 2.2. Different tissues are characterized by different attenuation coefficients which enables a visualization of the patients morphology. By shifting the patient in a plane perpendicular to plane containing the X-ray source and the detector array, a two-dimensional intensity profile can be determined. The resulting two-dimensional image is called a *radiography*. Highly absorbing structures, such as bones, appear as high intensity areas in X-ray images (since the image is digitally inverted due to historical reasons), whereas low absorbing structures, such as soft-tissue, result in low intensity values. Thus, X-ray imaging is characterized by a low soft tissue contrast, but a high image contrast for high density objects. This is exploited for various imaging techniques, such as e.g. angiography, where a highly absorbing contrast agent is injected to enhance certain features as e.g. vessels. In a context in which real-time moving images are acquired, the term *fluoroscopy* is used instead of radiography, which refers to fixed still images.

Computed Tomography

In the early 1970s, the development of Computed Tomography (CT) opened up various new possibilities in the field of diagnostic radiology, since it enabled the generation of three-dimensional cross-sectional images of the body and thus the visualization of superimposing objects which could not be distinguished in 2D representations such as radiographies [22].

For the calculation of a three-dimensional CT scan, multiple two-dimensional X-ray projections are acquired from different angles whilst rotating the X-ray tube and the detector array around the patient. In modern generation CT scanners, this is done by placing the patient in a detector array which is arranged as an outer circle and rotating the X-ray source in a spiral around the patient as shown in Figure 2.3. By applying a geometric reconstruction algorithm to these projections, such as i.e. a filtered back projection [23] or iterative reconstruction [24], a three-dimensional volume of the patient's anatomy is calculated from the measured intensity profiles. The grey values in a volume element (voxel) of a CT scan correspond to the mean attenuation coefficient of X-rays due

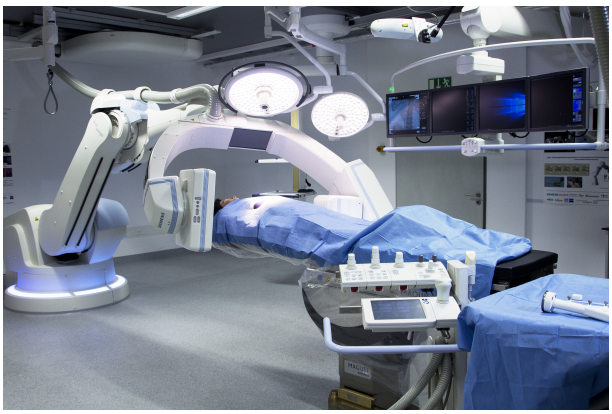


Figure 2.4: Artis Zeego C-arm system (Siemens Healthineers, Forchheim, Erlangen, Germany) implemented in the intervention room of the M^{2}OLIE research campus (image courtesy by Vanessa Stachel, Fraunhofer IPA).

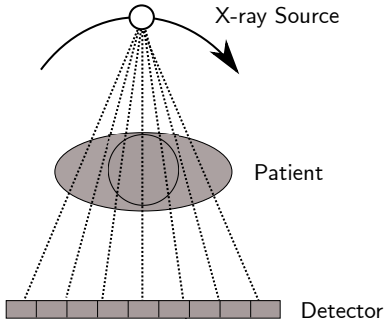


Figure 2.5: Setup of a Cone Beam CT.

to the absorption caused by the tissue covered in this voxel. To assure a standardized comparison between different CT images, the image intensity is measured in terms of *Houndsfield Units*. These units are defined such that the value of water corresponds to zero and can be estimated using the following relation

$$HU = \frac{\mu_{tissue} - \mu_{H_2O}}{\mu_{H_2O}} \cdot 1000. \tag{2.2}$$

The employment of Hounsfield Units allows to estimate electron densities for dose calculations. Due to their tomographic nature, CT scans enable the acquisition of detailed images of internal organs. The development of widened fan-beam X-ray sources as well as multi-array detectors corresponding to an increased width of the detector ring, allowed for a simultaneous acquisition of multiple images in scan plane during a single rotation. Since this so-called Multi-Slice CT (MSCT) grants a fast image acquisition in the range of a few seconds with a high spatial resolution. It can be used to record dynamic processes, such as the beating of the heart, by acquiring several CTs over time resulting in a 4D CT scan [25].

In general, CT is a very precise imaging technology, which offers a high speed of acquisition, and therefore a low risk of distortions due motion during the scan. Moreover, CT data usually provides a high spatial resolution and in comparison to other imaging modalities a pronounced dense tissue contrast. On the other hand, CT scans yield a low soft tissue contrast and the biggest disadvantage is the exposure of the patient to ionized radiation, which can cause damage to the patients DNA. Therefore a large field of research focuses on the development of image acquisition and reconstruction techniques which require less X-ray projections [26] and thus, less radiation dose.

Cone-Beam CT

Conventional CT devices do not support the acquisition of 2D projections or fluoroscopies which can pose a problem for interventional procedures that rely on real-time moving images to visualize e.g. vessels by observing the spreading of a contrast agent [27]. In addition, conventional CT devices

offer only a limited patient access and working space for the interventionalist. Therefore, open C-arm systems equipped with a Cone Beam CT source and a flat panel detector as exemplary shown in Figure 2.4 represent an important alternative, since they allow the acquisition of fluoroscopies as well as 3D volumes in a single orbit around the stationary positioned patient.

One of the main differences to a conventional CT is the employment of an X-ray cone beam geometry instead of a fan-shaped X-ray beam. Rather than spiraling the X-ray source around the patient by shifting of the patient table, the C-arm rotates the X-ray source and the detector around the stationary patient to acquire a three-dimensional image volume (shown in Figure 2.5). Since this enables the acquisition of a three-dimensional image volume with a single orbit, the acquisition time is reduced compared to a conventional CT that requires a spiral motion. Moreover, this acquisition leads to a reduction of radiation dose for the patient compared to the image acquisition using a conventional CT. However, the use of the cone beam geometry also results in a higher complexity concerning the image reconstruction to generate cross-sectional views. Due to the geometry of the image acquisition process, CBCTs are generally characterized by a circular shaped field of view in the image plane perpendicular to the rotation axis of the system and hexagonal shape in the planes parallel to this axis [28]. This limitation of the field of view is a major difference to other tomography techniques.

Another difference to conventional CT is the use of flat panel detectors instad of detector rings. On one hand, the employment of flat panel detectors offers the potential to use smaller pixel sizes which results in a high spatial resolution of the acquired image volumes. On the other hand, smaller detector pixel sizes lead to an increased level of Poisson noise due to a smaller number of photons that can be registered [29]. Moreover, the employment of a cone beam geometry instead of a fan beam combined with a flat panel detector results in an significant increase of scattered radiation in a CBCT [30]. The combination of both factors leads to a decreased image quality of a CBCT in comparison with a MSCT [31, 32]. Since C-arm systems are favorable to be used in an interventional environment as means for image-guided procedures, the image quality is often additionally affected by artefacts due to highly absorbing interventional tools. Another major difference in a CBCT compared to a CT is the dependence of the image value of a voxel of an organ on its position in the image volume [33, 34]. Thus, the values of a CBCT do not correspond to the HU values for similar structures in a conventional CT. A summary of typical image artefacts in CBCT data is given in [28].

2.2.2 Magnetic Resonance Imaging

As an alternative to X-ray imaging, research on Magnetic Resonance Imaging (MRI) started in the early 1970s and the first MRI prototypes were tested in the 1980s [35]. MRI represents a tomography imaging technique which relies on the varying magnetic properties of atoms to produce an image. Thus, the imaging process does not involve the use of ionizing radiation which is associated with potential harmful effects and is therefore considered a non-invasive imaging method. Since water represents one of the main components of the human body, mainly the electromagnetic effects of the nuclei of hydrogen atoms, thus single protons, are used for imaging. In general, atoms with an odd number of protons or neutrons possess a nuclear angular momentum. These nuclei can be

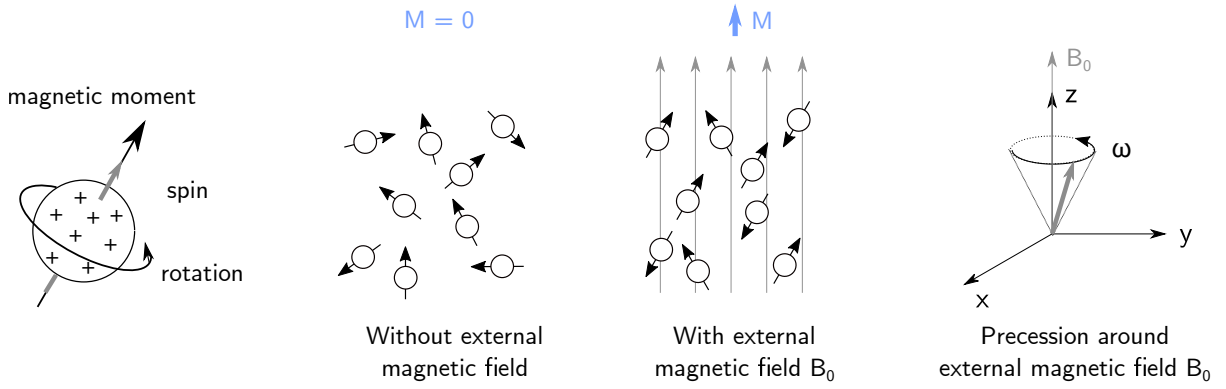


Figure 2.6: Alignment of magnetic moments with and without external magnetic field B_0. Without an external magnetic field, the spins are randomly aligned. With B_0 the spin of a hydrogen nucleus aligns either parallel or antiparallel to the external field and start to precess around its direction.

visualized as spinning charged spheres which possess a magnetic momentum as shown in Figure 2.6, referred to as spin. In the absence of an external magnetic field, these spins are oriented in random directions due to thermal motion and the summarized macroscopic magnetic moment M_0 equals zero. The quantum model restricts the spin of a proton to align either parallel or anti-parallel to an external magnetic field B_0 whereas the parallel alignment corresponds to a low-energy state and the anti-parallel alignment to a high-energy state. This results in an excess of parallel aligned spins creating a macroscopic magnetic moment $M_0 > 0$ when an external magnetic field is applied.

The fact, that the nucleus has an angular momentum and therefore is subjected to momentum conservation is leading to a precession around the axis of B_0 (see figure 2.6). The frequency of this precession is called Lamor frequency ω_0 and defined as

$$\omega_0 = \gamma \cdot B_0. \tag{2.3}$$

The factor γ refers to the gyromagnetic ratio, a known constant for each type of atom (for hydrogen: $\gamma/2\pi = 42.58\,\mathrm{MHz/T}$). Conventionally, the direction of the applied magnetic field is defined as z-axis. Under this assumption, the transverse component $M_{0,xy}$ defined as the projection of M_0 in the x-y-plane equals zero in equilibrium, since all contributions to $M_{0,xy}$ are dephased.

To obtain an MR signal, a radio frequency (RF) pulse is used to generate a secondary oscillating magnetic field B_1 which is applied perpendicular to B_0 using the resonance frequency matching ω_0. The spin precesses around the field direction of the superimposed magnetic field, causing the nuclear spin to spiral into a higher energetic state. Depending on the time that the oscillating field is active and its magnitude, the nuclear spin orientation is directed away from the z-axis, resulting in a transverse component $M_{0,xy} > 0$. If M_0 is e.g. flipped completely to the x-y-plane, the corresponding radio frequency pulse is referred to as 90° pulse, a 180° pulse inverts the bulk magnetization. The RF excitement is followed by an exponential relaxation, during which the system is restoring its equilibrium state. It is possible to measure $M_{0,xy}$ during these processes since it induces an electric voltage in another radio frequency coil. This signal builds the basis for the image formation in MRI.

However the relaxation process is complex, due to the superposition of two independent effects:

longitudinal and transverse relaxation. The longitudinal relaxation refers to the increase of the longitudinal magnetization (parallel to B_0) caused by an energy exchange between the spin system and the surrounding thermal reservoir referred to as 'lattice' (*spin-lattice relaxation*):

$$M_{0,z}(t) = M_0 \left(1 - e^{-t/T_1}\right). \tag{2.4}$$

Here, T_1 refers to the time constant of this relaxation process and t to the time since the application of the RF pulse. For $t \to \infty$, the magnitude of the longitudinal magnetization approaches the magnitude of the initial magnetization $M_{0,z} \to M_0$. The transverse relaxation describes the decrease of the transverse magnetization component $M_{0,xy}$ due to interactions with neighboring atoms (*spin-spin relaxation*) with time constant T_2:

$$M_{0,xy} = M_{0,xy} e^{-t/T_2}. \tag{2.5}$$

Thus, for $t \to \infty$, the transverse magnetization approaches zero $M_{0,xy} \to 0$. Generally $T_2 \leq T_1$. However, the transverse relaxation is not only influenced by spin-spin interactions, but also by inhomogenities in the local magnetic field, which is taken into account by defining the effective transverse relaxation time T_2^*.

T_1 and T_2 represent important quantities in MRI since they build the source of the image contrast. To obtain high quality relaxation signals, spin echo sequences are applied which describe a certain sequence of RF pulses to readout T_1 and T_2. A spin echo sequence consists of a 90° RF pulse followed by a 180° pulse. The 90° RF pulse tips the magnetization into the x-y-plane. $M_{0,xy}$ can then be measured to define T_2^*, whereas the signal directly after a high frequency pulse is called Free Induction Decay. The spins start dephasing in the x-y-plane. By applying a 180° pulse, the spins flip in the opposite direction in the transversal plane and start to refocus resulting in a pronounced transverse magnetization signal, the so-called *spin echo*. Echo pulse sequences are characterized by the echo time T_E which denotes the time between the 90° pulse and the spin echo signal. When repetitive spin echo sequences are used, the time between the application is called repetition time T_R. It is possible to measure T_1 by applying several 90° pulses with short T_R, since it flips the non-relaxed longitudinal magnetization in the x-y-plane where its signal can be measured. T_2 can be measured by repetitively applying spin echo sequences measuring the amplitude decay of the spin echo signals.

To spatially encode the source of measured magnetization signal, a magnetic gradient field is superimposed to the constant external field B_0 between the excitement with an RF pulse and the signal measurement. As a result, the effective Larmor frequency of the precession is depended on the local magnetic field $\omega_0(z) = \gamma \cdot (B_0 + B(z))$, and the previously aligned spins are artificially dephased. Thus, all spins precess in the same frequency but different phases, however, spins in the same row perpendicular to the gradient direction have the same phase. This is called phase encoding and translates to the phase of the measurement signal. This procedure can be extended by frequency encoding, which is also based on the use of a magnetic gradient field perpendicular to the gradient field for phase encoding. In contrast to phase encoding, this field is continuously applied during the signal measurement, causing the previously dephased spins to rotate with different

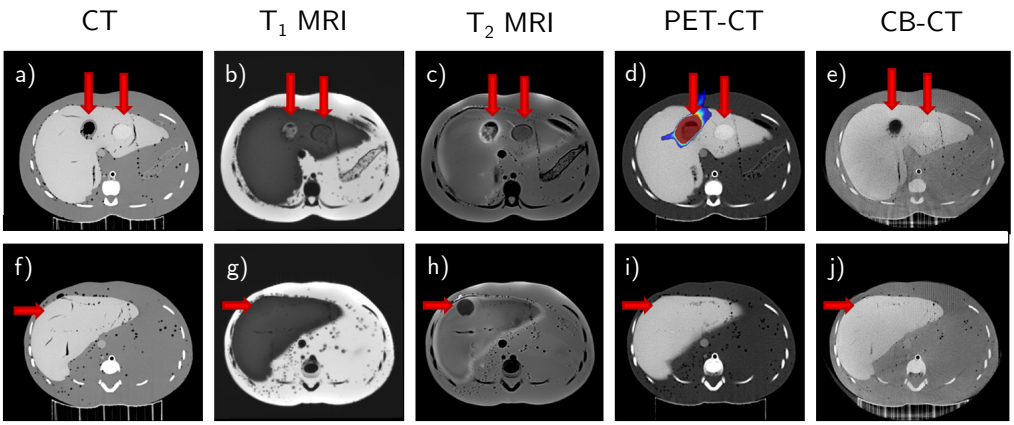


Figure 2.7: Axial slices through an antrophomorphic phantom extracted at two different axial positions. The phantom is specifically designed to demonstrate the advantages of multimodal imaging, since the composition of the artificial lesions is chosen in such a way that the visibility varies in different image modalities. The phantom is measured under clinically relevant conditions using the different imaging devices: a) + f) Multislice CT; b) + g) T1-weighted MRI; c) + h) T2-weighted MRI; d) + i) PET-CT; e) + j) Cone Beam CT. The red arrows indicate the positions of the artificial lesions.

frequencies depending on the spatial position. By applying a reconstruction based on a Fourier transform and a complete sampling of the data matrix containing the raw MRI data, the so-called k-space, it is possible to generate a spatially resolved image based on the local magnitude of the transverse magnetization. This magnitude is proportional to the proton density of the tissue, and thus it is possible to deduce a proton weighted image. By varying T_R during the application of a series of 90° pulses, a T_1-weighted image can be calculated, whereas the signal decreases with an increasing T_1. Choosing long T_R and T_E when applying spin echo sequences allows to calculate T_2-weighted images, in which the signal is proportional to T_2. Thus, varying the time parameters during a measurement leads to different image contrasts.

This represents one of the main advantages of MRI, since it is possible to generate images with different contrasts using the same imaging device. Moreover, MRI offers a very pronounced soft tissue contrast, since the image generation is based on the spin of hydrogen nuclei, and even allows to perform functional imaging [36]. However, the application of measurement sequences takes up to several minutes, making it time-consuming and the images prone to motion artifacts due to patient movements. Additionally inhomogeneities in the magnetic field or effects such as chemical shift can lead to small distortions of the morphology [37, 38].

In summary, CT generally offers a stable spatial resolution, fast acquisition times and a pronounced contrast for high-density structures such as bones, whereas MRI as a non-invasive technique is favorable for the imaging of soft tissue structures. Thus, it is often essential to employ multimodal imaging for diagnosis and treatment planning, to integrate all necessary information. The advantages of multimodal imaging are demonstrated by the measurement of an antrophomorphic phantom (designed within the research campus M^2OLIE) containing an artificial liver with three lesions: Figure 2.7 shows that some lesions are only visible in one modality, thus demonstrating the need for multimodal imaging. The material composition of the lesions has been specifically chosen to yield different contrast in different imaging devices.

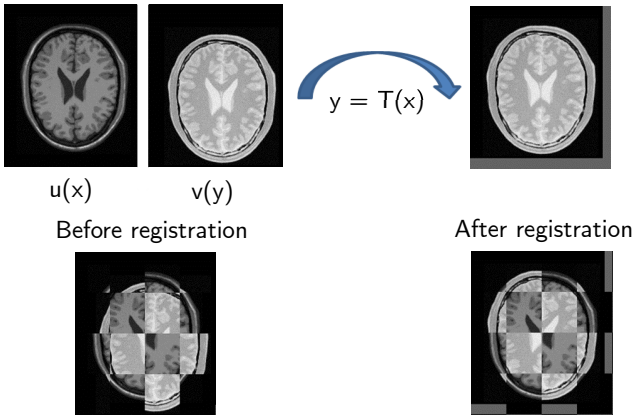


Figure 2.8: Optimization process of a geometrical transformation T so that corresponding features in the images overlap.

2.3 Fundamentals of Image Registration

The main challenge in the context of multimodal imaging of a patient is the fact that images acquired on different devices do not occupy the same physical space. Moreover, the organs of the patient are subjected to deformation due to different positioning of the patient in different devices, organ motion or image artefacts, impeding the comparison between different scans. Here, image registration plays a major role to geometrically transform the images and enable the fusion of different modalities in to a multimodal data set.

Image registration generally describes the process of determining a geometrical transformation to align two images of the same object or scene and transfer them into the same coordinate system so that identical features overlap as shown in Figure 2.8. There are various situations in which this is necessary, including scenarios in which images of an object are acquired from different angles, with different devices or at different points in time. Depending on the intended application, image registration is used to either determine the geometrically transformed image data or the transformation parameters. Image registration builds the basis to compare images of the same object and is therefore important in many fields of application, ranging from computer vision and pattern recognition, to medical image processing or even geosciences, when it comes to compare data from satellites [39, 40]. In a clinical context, image registration is e.g. necessary to align images acquired on different imaging devices or for treatment monitoring of diseases over time.

In a registration scenario, one of the images is chosen as reference image which is stationary and is therefore called *target*, *reference* or *fixed image*. The other image is referred to as *source* or *moving image*, since it is the image which is going to be geometrically transformed to align with the target image. Image registration then aims to find a reasonable transformation so that the transformed version of the source image is similar to the reference image.

Most image registration methods are considered as iterative optimization processes which aim to find the optimal parameters of the geometric transformation to spatially align two images so that the similarity between both images increases. This process is typically composed of four basic elements as shown in Figure 2.9, namely:

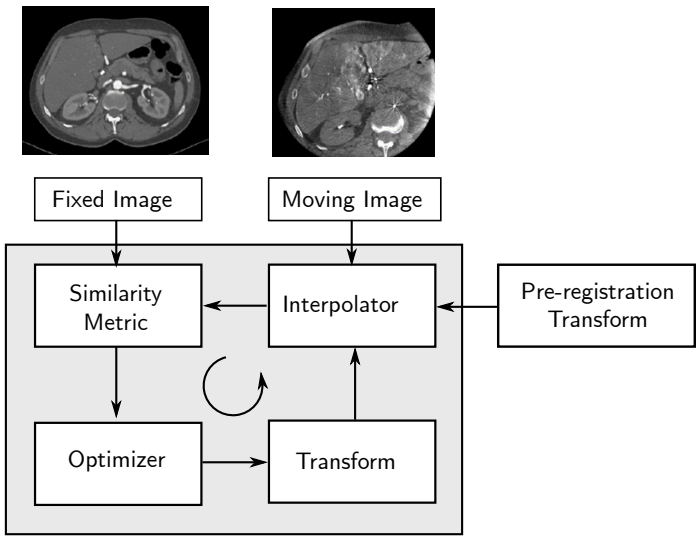


Figure 2.9: Visualization of the different components of a registration method and their interplay.

- A **transform** to geometrically distort the source image so that it aligns with the target image. The type of transformation also defines the number of transformation parameters which are optimized during the registration process.
- An **interpolator** to resample voxel intensities to the new coordinate system according to the geometric transformation.
- A **similarity metric** which measures the (dis)similarity of images for the different geometric alignments during the registration process based on their intensities, geometric features or higher-level information, such es e.g. segmentation labels.
- An **optimizer** which aims to find the optimum, i.e. minimum or maximum, of an energy function consisting of a similarity term or regularization penalty.

These components are often extended by a pre-registration transformation (or initialization) which prealigns the target and source image in terms of a higher image similarity. Initializations can be realized by overlaying the geometric image centers, by overlaying the intensity based centers of mass, by identifying and matching corresponding points [41] or features [42] in the images or by extracting and aligning segmentation masks [43].

All of these steps play a major role for the outcome of a registration method, and will therefore be explained in more detail in the following sections 2.3.2 – 2.3.5. Image registration approaches display a high diversity concerning the processed image information and structure of the methods. Therefore, a short overview of the different types of image registration algorithms will be given in the next section.

2.3.1 Classification of Image Registration Methods

There exist various different registration approaches, which can be sorted into different categories depending on various characteristics. According to Maintz *et al.* [4] these characteristics include:

- **Image dimensionality**: *2D-2D, 3D-3D, 2D-3D or time series*
  Registration methods are divided according to the spatial dimensionality of the input images for which it is distinguished between 2D-2D, 3D-3D and 2D-3D registration methods. Most medical registration methods focus on 3D-3D registration methods to register tomographic data sets or 2D-2D registration methods to register slices of tomographic data. 2D-3D methods are e.g. used for the registration of a preoperative 3D image to an intra-operative projective image. In clinical applications, medical images are sometimes also acquired over time intervals to examine medical processes such as e.g. tumor growth, leading to the acquisition of 4D data sets. High image dimensionality corresponds to a high information density which needs to be processed by the registration algorithm and therefore generally increases the registration complexity as well as the computation time.

- **Nature of transformation**: *rigid, affine, deformable*
  In general, it is distinguished between rigid and non-rigid transformations. Rigid transforms are limited to translation and rotation and extended to affine methods by including scaling and shearing whereas non-rigid or deformable methods result in the generation of complex deformation fields. The type of the transformation determines the number of transform parameters which are optimized during registration and is therefore directly related to the complexity of the registration task. An extended overview over the different transformation models is given in section 2.3.2.

- **Domain of transformation**: *global vs. local*
  The domain of transformation describes the image area on which the transformation is applied. A *global* transformation is applied on the entire image whereas a *local* transformation warps only a subsection, a so-called Region Of Interest (ROI). Registration methods are most commonly employed on a global basis, i.e. a global geometric transformation is applied to warp the source image to the target image.

- **Nature of registration basis**: *extrinsic vs. intrinsic*
  Extrinsic registration methods are based on the detection of foreign objects (markers), which are introduced to the image space and well visible in the image data. This allows a fast and easy registration by aligning the artificial object whereas the main drawbacks are the often invasive character of the object and the fact that provisions must be made before the image acquisition. Moreover, since extrinsic methods rely on the alignment of external objects, no patient information is included making it an insufficient method for tasks such as soft organ alignment.

  In contrast, intrinsic methods rely on patient generated image content only. This content can be represented by salient image points (landmarks), by segmented binary structures such as object surfaces (segmentation based) or by the intensity distribution of the grey values in the image (voxel property based). In medical image processing, landmarks are points in the patients morphology which can be accurately detected and located. By identifying and matching these points on two data sets, it is possible to estimate a geometric transformation to align the images. In many applications, these points are interactively identified by a user which makes these approaches unsuitable for daily clinical routine. However, in some

cases where the landmarks are based on geometrical properties (corners, local curvature), an automatized identification is possible. A high density of such landmarks allows for an accurate registration including deformable tasks. Segmentation based methods extend this approach by identifying and aligning surfaces or volumes in different image data sets. In contrast, voxel property based methods rely directly on the image grey values without including prior knowledge as landmarks or segmentation based methods. To establish a similarity measure between the images, the grey value distributions in the images are analyzed via correlation metric, Fourier properties or other means of structural analysis. A brief overview over some of the most commonly employed intensity-based and approaches for feature-based metrics is given in chapter 2.3.4.

- **Interaction**: *interactive, semi-automatic, automatic*
  Concerning registration methods, three levels of user interaction are distinguished. Interactive processes include all methods in which a user performs the registration himself, but is assisted by software giving him visual or numerical feedback. Semi-automatic methods require either a user performed initialization of the image alignment or user generated input data such as e.g. segmentation labels or user feedback in form of a rejection of acceptance of the suggested registration hypotheses. Most approaches aim to realize an automatic registration method, which requires only the image data and limited information on the image acquisition by the user. Automatic methods provide a high comfort for the user making it suitable for daily clinical practice, however the method has to be very robust to limit potentially false registration results.

- **Optimization procedure**: *direct, indirect, multi-stage*
  Since image registration represents an iterative optimization procedure, the results highly depend on the choice of optimization. The required transform parameters can either be computed directly from the available data or searched for by finding the optimum of a cost function. A direct computation is often only feasible for sparse data (as e.g. shown in [44] or [45]). To search the optimal parameter setting a quasi-convex mathematical cost function has to be defined depending on the transformation parameters to quantify the similarity between the images. The optimization then aims to identify the optimum of this function. A summary of most commonly used optimization techniques used in image registration is given in chapter 2.3.5. Additional approaches to accelerate convergence of the optimization process include multi-resolution approaches during which the spatial image resolution is altered from coarse to fine during registration or multi-stage approaches, during which a rigid registration to roughly align the images is ensued by a deformable registration for fine alignment.

- **Involved image modalities**: *monomodal, multimodal, modality to model, patient to model*
  Monomodal registration methods refer to the alignment of two or more images which were acquired using the same imaging device whereas multimodal registration involves the mapping of input images acquired on different devices, e.g. MRT and CT. The latter results in a highly increased difficulty for the registration method, since different imaging modalities are based on different physical principles and therefore often display dissimilar object structures as discussed in chapter 2.2. Thus, it represents a challenge to establish a relation between the

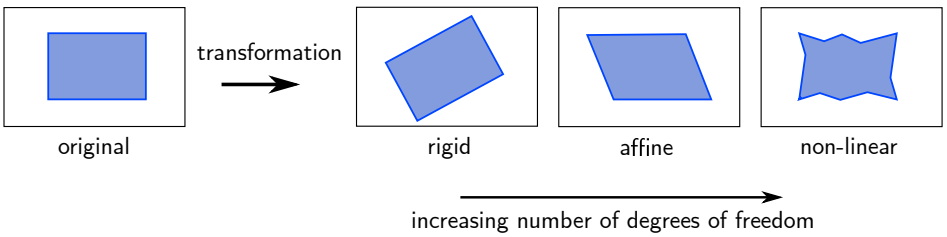


Figure 2.10: Visualization of 2D transformations using the example of a rectangle.

input images to estimate and optimize an image similarity measure. Additional registration scenarios include the registration of an image modality to a mathematical model or the patient himself.

- **Registration subject**: *intra-subject, inter-subject, atlas*
  Intra-subject registration refers to tasks involving the image data of a single patient whereas inter-subject registration is accomplished when the data of different patients is registered. Atlas registration describes the task to register the image data of a patient to a constructed image from an image formation database, such as binary segmentation masks.

An extensive summary of these criteria is given in [4] and [46]. These categories show that there exists a high diversity concerning image registration methods with profound differences in complexity and application possibilities. In this thesis, the focus is set to intra-subject mono- and multimodal 3D-3D registration techniques which provide a wide range of applications in clinical practice.

2.3.2 Geometric Transformation

A geometric transformation maps points from one image space to another. The choice of a geometric transformation model highly depends on the nature of the data to be registered and can be crucial for the success of a registration, since it defines the possibilities how an image can be warped. Some transformation types result in global transforms which are applied to the entire image whereas other transformation types yield local transforms which are only applied to a sub-region of the image and are therefore useful to correct small organ movement or deformations. In general, it is distinguished between linear and non-linear transformation types whereas the term 'linear' refers to the function which is used for the mapping of one vector space to the other. Both types will be briefly explained in the following. A visualization of the most commonly used geometric transformations is shown in Figure 2.10.

Rigid Transformation

The simplest type of linear transformations are rigid transformations. Rigid transformations comprise geometric transformations which preserve the Euclidean distance between every pair of points in an image, which means any object will keep its shape and size after the application of a rigid transformation. Rigid transformations include translations and rotations as well as their combinations. They are very useful for the registration of rigid structures, such as e.g. the skull, or an

initial alignment of two images, but fail to compensate non-linear motions or organ deformations. However, due to its limited degrees of freedom (DOF), the employment of a rigid transformation allows for a fast and simple image registration since less parameters have to be optimized.

Mathematically, a rigid transformation $T : \mathbb{R}^3 \mapsto \mathbb{R}^3$ of the 3D point $\mathbf{x} = (x, y, z)$ can be formulated as

$$T_{rigid}(\mathbf{x}) = R\mathbf{x} + \mathbf{t} \tag{2.6}$$

Here, $\mathbf{t} = (t_x, t_y, t_z)$ denotes the translation vector and R the rotation matrix. There are different approaches to describe an arbitrary rotation either including Euler angles α_1, α_2, α_3 or e.g. quaternions. For 3D data, a rigid transformation is defined by the three components of the translation vector $\mathbf{t}$ and three parameters describing the rotation angles. This results in six DOF in 3D and four in 2D.

Affine Transformation

Affine transformations represent the simplest type of non-rigid transformations. They extend rigid transformations by shearing and scaling. Angles between lines or distances between points are no longer preserved, although the ratio of distances between points on a straight line are maintained as in the original image. Thus, they are also considered as type of linear transformation. A complex affine transformation T_{affine} can be represented by a sequence of basic transforms. This if often described using homogenous coordinates which are a concept that stems from the mathematical field of projective geometry. They represent an extension of standart three-dimensional vectors and allow the simplification of various transforms and their computation. In homogenous coordinates, the sequence of basis transform to generate an affine transform can be expressed as a matrix multiplication such that

$$T_{affine} = T_{translation} \cdot T_{rotation} \cdot T_{shear} \cdot T_{scaling} = \begin{pmatrix} m_1 & m_2 & m_3 & m_{10} \\ m_4 & m_5 & m_6 & m_{11} \\ m_7 & m_8 & m_9 & m_{12} \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{2.7}$$

Here, m_{10} to m_{12} denote the translation parameters whereas all other parameters define scaling, shear and rotation. This results in 12 DOF in a 3D scenario, and six in a 2D scenario.

Nonlinear Transformation

The previously presented linear transformations mainly capture global image motion, but are not sensitive to model local structure deformations. This limitation is overcome by employing non-linear, or deformable, transformations which represent the most complex type of geometric transformations. They do not preserve straightness of lines nor parallelism, making them suitable to model complex tissue deformations. This requires a high number of DOF and is generally a sophisticated task.

There exist different approaches to model non-linear deformations including non-parametric *guided deformations* and multi-parametric *basis function transformations*. Guided deformations are based on physical models which limit the range of potential transformation parameters due to the assumption that these deformations have to follow physical principles. As the name suggests, basis function transformations rely on a set of basis functions to model the transformation. Both types are used in image registration. Parametric registration in this context simply implies that the transformation is indirectly optimized by optimizing a parametric model which governs the transformation, whereas non-parametric registration means that the transformation is directly optimized. Thus, the distinction between parametric and non-parametric registration does not describe the transformation, as rather the type of regularization or parametrization which is used during the registration process.

The introduction of a regularization is necessary, since non-linear image registration is generally considered as ill-posed problem. The estimated transformation during registration is required be plausible in the sense that the determined deformation field should be generally smooth so that it displays no singularities, tearing or folding. To avoid this, it is necessary to apply constraints on the DOF of the deformation in form of a regularization. This can be done implicitly or explicitly. Implicit regularization is achieved by parametrization of the deformation field with smooth functions. Explicit regularization can be realized in form of constraints that the solution must satisfy to achieve a successful registration or as additional penalty term in the energy function that is optimized during registration.

For both non-linear transformation types, the actual transformation in 3D is represented by a deformation field $\mathbf{u} : \mathbb{R}^3 \mapsto \mathbb{R}^3$, which corresponds to a displacement field such that

$$T_{non-linear}(x) = x + \mathbf{u}(x) \tag{2.8}$$

Nonlinear transformations based on physical models take into account tissue characteristics, such as e.g. elasticity, flow or diffusion. Following the idea of an elastic model, the objects in an image are modeled as elastic solids which are affected by internal elastic forces as well as external forces driven by similarity measures which try to deform the objects. In this scenario the source image is deformed until external and internal forces reach an equilibrium [47, 48, 49]. Another type of elastic transformation is based on finite element models which divide the input image into cells with certain tissue properties [50, 51]. Flow based transformations use physical fluid flow models in which the moving image is represented as a viscous fluid and the deformations can be described using formulas such as the Navier-Stokes-equation [52, 53, 54]. In general, they allow larger deformations than elastic model transformations. The most popular registration approach using diffusion models is presented by Thirion *et al.* [55]. Here, the main idea is that object boundaries in an image are considered as semi-permeable membranes through which the other image can diffuse by action of effectors ("demons") situated within these interfaces.

In case of deformation models which rely on basis function models, the deformation is modeled by defining a regular grid of control points which can be moved individually in the direction that leads to an optimization of a similarity measure. The density of the grid points is proportional to the sensitivity of these methods to local deformations, but also to the computational cost of these approaches, since displacement vectors for each point have to be estimated. Thus, by varying

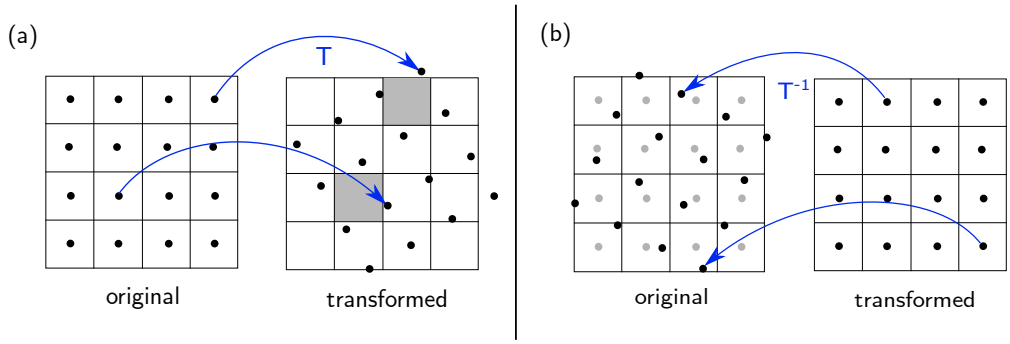


Figure 2.11: Forward and backward image warping using a transformation T and its inverse transformation T^{-1}, respectively. Forward warping (a) can cause non-defined pixel as marked in grey. Backward warping (b) allows to obtain intensity values for locations that do not coincide with pixel coordinates from the reference image using interpolation.

the number of grid points from low to high, a transition from the modeling of global to local deformations can be realized. These types of transformation models have in common that they rely on a combination of basis functions to determine the deformation field in between grid points. The most popular interpolation method relies on cubic (B-)splines [56, 57, 58] which will be further explained in the following chapter. Thus, the task of registration for these methods corresponds to the problem of finding a set of coefficients for the basis functions to optimize the image alignment in terms of a similarity measure. In contrast to physical models, transformations relying on basis functions employ much fewer DOF and are therefore favorable for image registration tasks. In [39], Sotiras *et al.* present an overview of deformable registration approaches in medical imaging discussing the different deformation models in more detail.

An important criterion for geometric transformations in terms of image registration is *diffeomorphism*. A diffeomorphic transformation is differentiable and invertible and thus, has a differentiable inverse transformation. This is important, since the transformation determined by a registration algorithm generally describes the transformation of the coordinate system of the source image to the coordinate system of the reference image. This is necessary to perform backward interpolation which is required to preserve image topology. Backward interpolation is preferred over forward warping since the latter can cause non-defined areas in the transformed source image as illustrated in Figure 2.11. This approach is often received as counter-intuitive, since e.g. the transformation obtained for the registration of a source image which is shifted to the right with respect to a target image will point to the left, and not to the right as expected. The basics of image interpolation will be explained in the following chapter 2.3.3.

2.3.3 Interpolator

Another basic component of a medical image registration method is the interpolator. Interpolation is necessary since the registration input consists of discrete pixelated (or in the context of 3D data voxelated) images which possibly differ in terms of spatial resolution and field of view (FOV). Due to this pixelated nature of an image, the intensity values at integer grid locations are known. However, the application of a geometric transform during the registration process possibly maps an intensity value to a sub-pixel position. Interpolation then uses known intensity values at integer

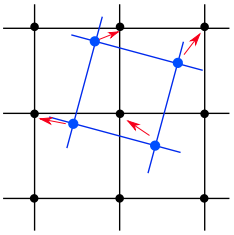


Figure 2.12: Nearest neighbor interpolation. The integer grid is displayed in black and the sub-pixel grid in blue.

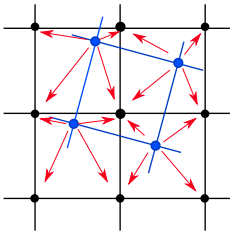


Figure 2.13: Linear interpolation. The integer grid is displayed in black and the sub-pixel grid in blue.

grid positions to estimate the intensity at these sub-pixel positions. The choice of interpolation method has an important impact on the performance of a registration method, since it affects the smoothness of the cost function and thus, the optimization process [59, 60, 61]. The most commonly used interpolation methods include nearest neighbor interpolation, linear and spline interpolation which will be briefly explained in the following. The explanation is based on the summary given by Forsberg in [62].

Nearest Neighbor Interpolation

Nearest neighbor interpolation assigns the value of the nearest sample point in the input image to the interpolated output pixel as shown in Figure 2.12, neglecting fractional contributions of other pixels. So basically, this methods relies on simply copying intensity values, not actually interpolating values. Let us assume that $\tilde{\Omega}$ represents the discrete domain of the image and Ω a continuous domain and $\tilde{x}$ and x are points in the respective domains. The nearest neighbor interpolation can then be defined as

$$I_{nearest}(\tilde{I}, x) = \tilde{I}([x]) \ , \tag{2.9}$$

whereas $[x]$ corresponds to the closest grid point $\tilde{x}$. This interpolation method is very simple, and thus, very fast, but lacks accuracy and often results in block structures for high magnifications of an image and significant aliasing effects along edges. Therefore, it is rarely used for image registration.

Linear Interpolation

A slightly more complicated interpolation method is *linear interpolation*. Under the assumption that intensity values change linearly from one pixel to the next, it estimates the final interpolated value by calculating the weighted average of the nearest neighbor pixels (see figure 2.13). For two dimensions this neighborhood corresponds to an area of 2×2 pixel, in three dimensions the area comprises $2 \times 2 \times 2$ voxel. Under the assumption that $\lfloor x \rfloor$ corresponds to the closest grid point smaller than $\tilde{x}$, d to the number of image dimensions and $\xi = x - \lfloor x \rfloor$, then the linear interpolation can be expressed as

$$I_{linear}(\tilde{I}, x) = \sum_{k \in \{0,1\}^d} \left[\tilde{I}(\lfloor x \rfloor + k) \prod_{l=1,\ldots,d} (\xi_l)^{k_l} (1 - \xi_l)^{(1-k_l)} \right] \ . \tag{2.10}$$

The interpolated pixel value then corresponds to a distance-based normalized linear combination of the four closest pixel intensity values. In comparison to the nearest neighbor interpolation, this

approach leads to a smoother interpolated image, but can lead to blurring effects. However, linear interpolation represents an appropriate trade-off between computation time and accuracy and is therefore often used for image registration.

Spline Interpolation

More sophisticated interpolation methods take even more than only adjacent voxel into consideration by fitting the values between the closest neighboring pixel using Lagrange polynomials, cubic splines, etc. Using the example of spline interpolation, a spline of degree $n \geq 1$ is a continuous piece-wise polynomial function of degree n of a real variable with continuous derivatives up to order $n-1$. One of the most commonly used family of spline functions are basis splines (B-splines) which can be derived by self-convolutions of a basis function

$$\beta^0(x) = \begin{cases} 1 & |x| < \frac{1}{2} \\ \frac{1}{2} & |x| = \frac{1}{2} \\ 0 & |x| > \frac{1}{2} \end{cases} \tag{2.11}$$

B-splines of the degree n are then iteratively defined using convolutions denoted with the operator symbol $*$:

$$\beta^n(x) = \underbrace{\beta^0 * \beta^0 * ... * \beta^0(x)}_{(n+1)\text{times}}. \tag{2.12}$$

Interpolation based on B-spline functions of degree $n = 0$ is similar to nearest neighbor interpolation and interpolation based on B-splines of degree $n = 1$ is identical to linear interpolation. Since cubic B-splines ($n = 3$) provide the best trade-off between performance and computation efficiency, they are most commonly used for image processing tasks and are given as

$$\beta^3(x) = \begin{cases} (2+x)^3 & -2 \leq x < -1 \\ -(3x+6)x^2 + 4 & -1 \leq x < 0 \\ (3x-6)x^2 + 4 & 0 \leq x < 1 \\ (2-x)^3 & 1 \leq x < 2 \end{cases} \tag{2.13}$$

Using a cubic B-spline interpolation, the intensity interpolation in one dimension at position x can be calculated as

$$I_{spline}(x) = \sum_{j=1}^{m} c_j \beta_j^3(x). \tag{2.14}$$

Here, m corresponds to the number of nearest neighbors, thus the number of available data points, c_j to the spline coefficients which need to be estimated for interpolation and the cubic B-splines fulfill the condition $\beta_j^3(x) = \beta^3(x-j)$.

The spline coefficients c_j are then determined using the constraint that $I_{spline}(\tilde{x}) = \tilde{I}(\tilde{x})$. Spline interpolations generally results in a high accuracy, however for high-resolution medical images linear interpolation is often sufficient and preferred due to its lower computational cost. An overview over the most common interpolation methods and their performance in the field of medical image processing is given by Deserno *et al.* [63].

2.3.4 Similarity Metrics

With regard to the typical components of iterative registration algorithms as shown in Figure 2.9, geometric transformations and interpolation methods have been discussed so far. A registration algorithm now aims to optimize the parameters of the transformation by defining and minimizing a cost function. Given a transformation T defined by a set of parameters θ and two input images I_t and I_s, this can be mathematically expressed as

$$\theta^* = \arg\min_{\theta}(d(I_t, I_s)) \ . \tag{2.15}$$

Here, d denotes a similarity measure which estimates and quantifies the (dis)similarity between two images. An ideal similarity measure corresponds to a smooth function which displays a global extreme at the position of optimal alignment of the two images. In general, it is distinguished between image-based similarity measures and feature based similarity measures. Depending on the utilized features, some similarity measures can be included in both types of similarity measures which will be further explained in the following.

Intensity-based Methods

Intensity-based similarity measures include all methods which rely on statistic correlations between the intensity distributions of the target and source image to estimate image similarity. This is based on the idea that similar structures in images share similar intensity value distributions and that correlations are maximized when these two images are perfectly aligned. The correlation metrics do not only rely on linear dependencies, but also include various relations such as e.g. entropy. There exist many different variants of intensity-based similarity measures, however, in this summary only the most commonly used metrics will be presented.

Sum of squared differences
A simple way to quantify image similarity between a reference image I_r and the source image I_s is to determine intensity differences for each voxel position in two images based on the assumption that corresponding structures in two images have similar intensity values. This approach is called *sum of squared differences* (SSD) and is defined as

$$SSD(I_r, I_s) = \frac{1}{N} \sum_{\mathbf{x} \in \Omega} (I_r(\mathbf{x}) - I_s(T(\mathbf{x})))^2 \ , \tag{2.16}$$

where Ω denotes the common image domain of both images, N the total number of pixels in this common domain and T the transformation function that maps a voxel at position $\mathbf{x}$ to its new position. The lower the SSD, the higher is the image similarity. Since SSD relies on the assumption that corresponding data points have corresponding grey values, its performance is highly sensitive to a small amount of pixel with large intensity difference. This can be especially important in case the metric should be applied to very noisy images. A modified version of SSD is the *sum of absolute*

differences (SAD) which eliminates the quadratic behavior and is given as

$$SAD(I_r, I_s) = \frac{1}{N} \sum_{\mathbf{x} \in \Omega} |I_r(\mathbf{x}) - I_s(T(\mathbf{x})|^2 \ . \tag{2.17}$$

In comparison to SSD, SAD is less sensitive to large intensity differences. However, both of these similarity measures are influenced by illumination changes leading to the requirement of intensity normalization of the input images. The metrics are most suitable for the similarity estimation of images with corresponding grey values.

Cross-correlation

Another concept to estimate image similarity relies on *cross correlation*, a method which is widely used in signal processing to determine the correlation between two signals. In image processing, cross-correlation assumes that there exists a linear correlation between the intensity values of corresponding structures in two images and is given as

$$CC(I_r, I_s) = \frac{1}{N} \sum_{x \in \Omega} I_r(\mathbf{x}) \cdot I_s(T(\mathbf{x})). \tag{2.18}$$

To increase its robustness to intensity and contrast changes, the correlation is normalized leading to the definition of a *normalized cross correlation* (NCC) metric given as

$$NCC(I_r, I_s) = \frac{\sum_{x \in \Omega}(I_r(\mathbf{x}) - \mu_r)(I_s(T(\mathbf{x})) - \mu_s)}{\sqrt{\sum_{x \in \Omega}(I_r(\mathbf{x}) - \mu_r)^2} \sqrt{\sum_{x \in \Omega}(I_s(T(\mathbf{x})) - \mu_s)^2}}. \tag{2.19}$$

Here, μ_r and μ_s correspond to the mean of the intensity values computed over the overlapping domain Ω in the reference and source image, respectively. The higher the numerical value of the NCC, the better is the alignment of the two images. In contrast to SSD or SAD, cross-correlation based similarity metrics do not rely on grey value correspondences but rather on a linear correlation of the grey values of corresponding data points. Thus, these metrics are robust to illumination differences and image noise. Moreover, they are easy to implement and have a low computational demand and are therefore widely used in medical image registration [64, 65].

Mutual information

The described similarity measures represent commonly used methods for monomodal image registration tasks. However, these metrics are not suitable for multimodal registration applications, since the assumption that similar structures display similar intensity values is not valid for the case of multimodal data as discussed in chapter 2.2. In general, different image modalities display different image contrasts and can even display structural differences due to the physical principles of the imaging technique. This makes the registration of multimodal image data highly challenging.

The most popular approach to establish a relation between the intensity value distributions of multimodal image data relies on statistical relations between two images in form of *mutual information* (MI) or its variants. The general idea of MI is to quantify the information shared between two images using the Shannon entropy H of a probability distribution which is estimated from the distribution of intensity values in the images. The basics of information theory using entropy are

explained by Shannon *et al.* [66]. In general, mutual information can be expressed as

$$MI(I_r, I_s) = H(I_r) - H(I_r|I_s) = H(I_r) + H(I_s) - H(I_r, I_s) \ , \tag{2.20}$$

where $H(I_r)$ and $H(I_s)$ denote the marginal Shannon entropy of the reference image and source image, respectively, and $H(I_r, I_s)$ denotes the joint entropy of both images. The entropy is used as a measure of the amount of information contained in an image, whereas the joint entropy refers to the cumulated amount of information in both images. In general, the entropy of an image A is given as

$$H(A) = -\sum_{a \in A} p_A(a) \log p_A(a) \ . \tag{2.21}$$

Here, $p_A(a)$ denotes the probability that a voxel in image A has the intensity a. The probability distribution of an image can be deduced from a histogram of the intensity values. The joint entropy of two images A and B is then defined as

$$H(A, B) = -\sum_{a \in A} \sum_{b \in B} p_{A,B}(a, b) \log p_{A,B}(a, b) \ . \tag{2.22}$$

In this case, the probability distribution function $p_{A,B}$ is derived from a normalized 2D histogram which depicts how often each grey value correspondence in the image A and B occurs. The crucial property is that $H(A, B) \leq H(A) + H(B)$. If both images are completely unrelated the joint entropy corresponds to the sum of individual entropies. The image alignment is improved when the joint entropy $H(A, B)$ is minimized. Studholme *et al.* [67, 68] as well as Collignon *et al.* [69] proposed joint entropy as similarity measure in image registration so that the entropy is minimized during registration. Almost simultaneously, Viola and Wells [70, 71] and Maes and Colligonon [72] proposed to use MI as similarity metric. Over the years, different derivatives of the classical MI metric were publishaed, including e.g. Mattest Mutual Information (AMMI) where the probability distribution is estimaned using Parzen histograms [73, 74].

Because of the statistical notion, MI is generally very sensitive to the amount of overlap between both images. If the overlap of both images decreases, the number of samples decreases which reduces the statistical power of the probability distribution. Moreover, it can be shown that for the case of misregistrations which typically lead to a decrease of the overlap, MI possibly increases. This may happen when the relative image areas of object and background even out and the sum of marginal entropies increases faster than the joint entropy [5]. Therefore, Studholme *et al.*[75] proposed normalized mutual information (NMI) which is defined as

$$NMI(I_r, I_s) = \frac{H(I_r) + H(I_s)}{H(I_r, I_s)} \tag{2.23}$$

NMI is generally a less sensitive variant of MI to changes in the image overlap. Since the image overlap constantly changes during a registration process due to the applied geometric transformation, NMI represents a suitable and robust similarity measure for registration purposes.

Due to the fact that MI does not rely on a linear relation between the intensity values of two images but on statistical relations, it is reasonable to be used as similarity measure for multimodal image

data and MI and its many derivatives remain the state-of-the-art for multimodal image registration. However, MI is computationally expensive which makes the algorithm slow and therefore not suitable for applications requiring fast or even real-time image registration. A survey on image registration techniques based on mutual information was published by Pluim *et al.* [5].

Feature-based Methods

In contrast to intensity-based methods which use the entire image information during the registration, feature-based methods rely on the extraction, analysis and matching of distinct features in the images which are to be registered. These features can be represented by geometrical features such as points (so-called "'landmarks"'), lines and surfaces or by morphological features, e.g. anatomical landmarks or fiducial markers. Image registration is then performed by optimizing a geometric transformation which reduces the distance between corresponding features in both images. These methods are mainly divided into landmark-based and surface-based methods.

For both types, feature-based registration methods require a preprocessing step of the image data to extract these features manually or automatically. Manual preprocessing in medical image registration is very time consuming and the results of the identification of salient points or surfaces highly depend on the skill and medical expertise of the user. Automatized feature-based registration methods typically consist of three steps: 1) keypoint detection and feature description, 2) feature matching and 3) image warping.

In general, a keypoint is a distinct point of interest which can be clearly identified in an image, such as e.g. previously described landmarks or as corners and edges. In image processing, a keypoint can be represented by a *descriptor* which is a vector containing important characteristics of this keypoint and its neighborhood. In an ideal case, the descriptor is robust against transformations, including changes of brightness or geometrical transformations. Two of the most popular methods for feature description are the Scale-invariant feature transform (SIFT) [76] or its derivative Speeded Up Robust Features (SURF) [77]. There exist many alternatives for feature description whereas the most common image registration approaches relying on feature descriptors will be further explained in chapter 3.2.

Once the features are extracted and analyzed, image registration often relies on a matching method which identifies corresponding features in both images. A simple method to perform feature matching is by estimating the distance between each pair of keypoint descriptors and return the pair with the smallest distance as matching keypoints. However, feature matching remains a challenging task and various methods exist aiming to reduce false or double matching [78]. After a successful matching of the image features, it is possible to estimate a deformation field by determining the spatial distance between corresponding keypoints. Feature-based registration is not typically considered an iterative optimization process, since the features are only extracted and matched once two align the images. However, for a coarse rigid image alignment, the feature matching step can be neglected and the registration can be e.g. performed using an *iterative closest point* algorithm which minimized the difference between two clouds of points [79]. Another iterative feature-based registration scenario is realized when the feature descriptor is determined globally over the entire image and

does not only consider local features. This represents a mixture of feature- and intensity-based registration and in this case, an image similarity measure is e.g. represented by estimating the similarity of the feature descriptors of the images to be registered or estimating the Euclidean distance of the two descriptors for each optimization iteration.

Depending on the utilized features, feature-based registration can be suitable for mono- and multimodal image registration. The accuracy of these registration methods highly depend on the number and the accurate identification of the features. Thus, feature-based methods require either user support or highly accurate feature extraction algorithms. However, recent advances using deep neural networks allow to increase speed and accuracy of these procedures and therefore benefit medical image registration.

2.3.5 Optimizer

The last component of an image registration algorithm is the optimizer which aims to optimize the parameters of the transformation component with respect to the similarity value provided by the metric component. Thus, the similarity measure represents the cost function in this optimization procedure. As discussed in chapter 2.3.2, the amount of parameters which need to be optimized highly depend on the DOF of the geometrical transformation with simple rigid transformations having a low number of DOF whereas nonlinear transformations have a large number of DOF. A common challenge in image registration tasks is that the cost function contains local minima to which the registration possibly converges which leads to a misalignment of the images. To prevent the method from getting stuck in such a local minimum, different strategies established. One of them is multi-resolution registration which consists of different stages of image-down-sampling followed by a registration which serves as an initialization for the next, less coarse resolution level.

Since there exist a multitude of optimization problems in different kinds of research fields, various different optimization methods have been developed. The most commonly used optimization methods for image registraion include gradient descent [80], Simplex [81], Gauss-Newton [82], Quasi-Newton [83], Powell's [84], stochastic and the Levenberg-Marquart method [85]. A comparison of different optimizers for medical image registration was published by Klein *et al.* in [86]. Since most experiments presented in this thesis were performed using gradient descent optimization, only this method will be explained in more detail.

Gradient descent is a method to optimize an objective function, or in case of image registration a cost function, $J(\theta)$ which is parametrized by a set of parameters $\theta \in \mathbb{R}^d$. This method updates the parameters in the opposite direction of the gradient of the objective function with respect to the parameters $\nabla J(\theta) = \frac{\partial J(\theta)}{\partial \theta}$. The step size with which the parameters are changed to reach the minimum during the update is called the learning rate η. Thus, the new parameters are determine as

$$\theta_{new} = \theta - \eta \frac{\partial J(\theta)}{\partial \theta} \ . \tag{2.24}$$

This can be visualized such that the optimzer follows the slope of the parametric surface of the objective function downhill until a valley is reached as shown in Figure 2.14.

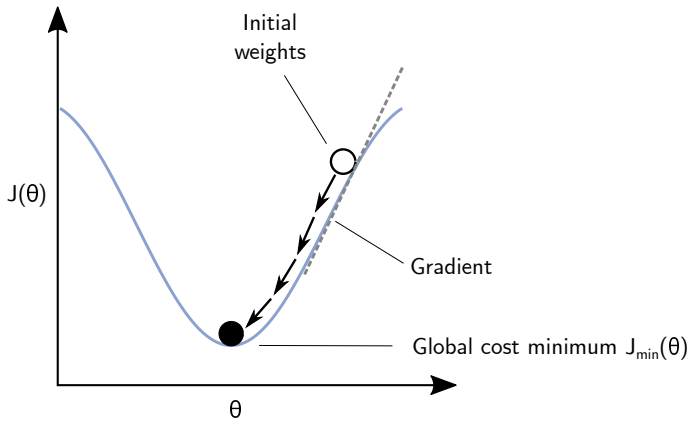


Figure 2.14: Optimization using a gradient descent method.

2.3.6 Challenges and Limitations

Image registration is still an active field of research which aims to develop novel fast and accurate registration methods, improve existing methods and broaden their range of applications in daily clinical routine. Despite ongoing efforts, the usage of image registration in clinical practice is often limited to rigid registrations of mono- and multimodal data [87]. One of the reasons why deformable registration methods are not in standard use is the uncertainty of these methods in terms of how physically plausible organ and tissue deformations can be corrected. Although there exist various approaches relying on physical models (as shown in chapter 2.3.2), image registration remains an ill-posed mathematical problem which requires regularization.

Moreover, more advanced registration methods are often restricted to clinical research due to their excessive computational cost which is either linked to high processing times or the demand of a sufficient IT-infrastructure. However, recent advances of graphics processing units (GPUs) highly improved computation times and helped to reduce this drawback. After years of theoretical model development, the introduction of GPUs in medical image processing even allowed the realization of novel registration techniques relying on deep neural networks, thus, opening a complete new field of research. The employment of neural networks allows for a fast (possibly real-time) deformable image registration, once the network is trained. The fundamentals of *Deep Learning* and the influence of neural networks on the field of medical image registration will be explained in the following chapter 3.3.

Nevertheless, advances using Deep Learning as well established methods suffer from the lack of golden standards and adequate validation methods for registration results. This is especially crucial for the training of neural networks which urgently requires ground truth data, but also hinders the optimization and improvement of classical method since there exists no standard method to evaluate registration accuracy which makes it difficult to compare registration performances of different methods for a certain application.

The work presented in this thesis aims to tackle this drawback by proposing an evaluation methodology for multimodal medical image registration of abdominal scans and using this framework to benchmark and optimize existing registration methods for this intended task.

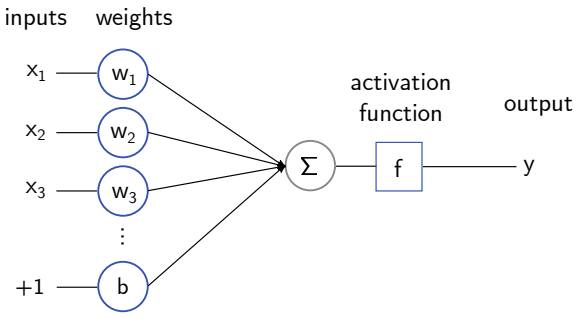


Figure 2.15: An artificial neuron.

2.4 Fundamentals of Deep Learning

Recent advances in the field of *Deep Learning* had tremendous impact on various fields of applications. One of them is medical image processing which has been highly benefiting from the employment of artificial neural networks in image detection and recognition, image segmentation, computer-aided diagnostics and also image registration [88, 89, 90]. This chapter aims to give a brief introduction to the fundamentals of Deep Learning including the basics of artificial neural networks and convolutional neural networks. Moreover, recent developments in image registration based on neural networks will be presented.

In general, Deep Learning (DL) refers to a sub-field of machine learning methods. In machine learning, it is generally distinguished between supervised, semi-supervised and unsupervised learning, depending on the input data that is used during the learning process. During supervised learning the algorithm receives both input and the desired output during the training process, as for example an image of an object and the corresponding label. This allows the algorithm to learn a mapping between those two inputs. In contrast, unsupervised learning does not provide the algorithm with the correct output. Here is the aim of the training process to find a structure in the given input. Semi-supervised learning refers to training processes during which only sparse prior information is provided, representing a mixture between both types.

These learning concepts also play a role in DL which is based on artificial neural networks. The general idea is to use algorithms which are inspired by the information processing and structure of the brain to mimic the learning process of a biological brain. To understand how the algorithm is able to extract an inner structure in the input data, it is essential to understand the information process of an artificial neural network.

2.4.1 Artificial Neural Networks

As artificial neural networks are inspired by the human biological nervous system, they consist of single neurons which are connected. In biology, a neuron collects inputs from all other neural cells to which it is connected and if the input reaches a certain threshold, it signals itself to connected neural cells. An artificial neuron, also referred to as perceptron, is the basic building block of an artificial neural network and can be seen as a simple model itself. It also receives inputs x_i from

multiple other neurons and computes the output as the sum of the weighted input values and a bias, which allows to add an offset to the data. To simulate the threshold, the result of this linear combination is then transformed via a nonlinear activation function f to get the final output of the neuron. Figure 2.15 displays the illustration of a single neuron. Mathematically, this can be expressed as

$$y = f\left(\sum_{i_1}^{m} w_i x_i + b\right) \ , \tag{2.25}$$

were w_i denotes the individual weights for each input x_i, the parameter b the bias and the function f the nonlinear activation function. By using the bias b, it is possible to adapt the activation threshold for the neuron. Combining and connecting several neurons results in an artificial neural network model.

The choice of *activation function* is important to model non-linear problems, since it should be continuously differentiable to enable the training of a neural network using back-propagation, a method which relies on gradient estimation and will be explained in more detail in the following section. A classic activation function is the logistic sigmoid function which is defined as

$$f(x) = \frac{1}{1 + e^{-x}} \ . \tag{2.26}$$

It is a smooth, non-linear function that is continuously differentiable and maps the input to values between [0,1]. However, a drawback of this function for DL is the fact that for large absolute values of x the gradient of the sigmoid function can become too small to be useful for a learning process. A generalization of the sigmoid function is represented by the softmax function which corresponds to a probability distribution and is given as

$$f(x) = \frac{e^{x_i}}{\sum_{i=1}^{K} e^{x_i}} \ . \tag{2.27}$$

This function takes a vector of K elements and outputs a vector of K values that range between $[0, 1]$ and sum up to 1. Therefore, the softmax function is often used for DL networks aiming to solve classification problems with more than 2 classes. Another activation function similar to the sigmoid function but symmetric over the origin is the hyperbolic tangent function:

$$f(x) = \tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}. \tag{2.28}$$

However, this function also has a vanishing gradient for large values x which can be a drawback for the training of a neural network using backpropagation.

The most widely used activation function is the rectifier (and its variants) [91, 92] which is given as

$$f(x) = max\left\{0, x\right\} \ . \tag{2.29}$$

A unit with this activation function is called rectified linear unit (ReLU) and is easy to optimize since the derivative is either 0 or a positive constant value which allows the network to converge quickly. A drawback of this function is the fact that inputs that approach zero or are negative

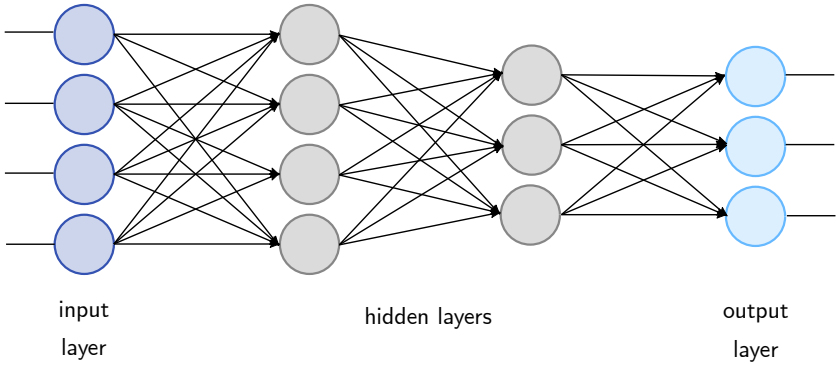


Figure 2.16: A fully connected multi-layer neural network.

result in a gradient equal to zero which hinders the performance of backpropagation.

In an artificial neural network, the neurons are arranged in layers, so that each layer receives input from the previous layer, applies weights and then signals to the next layer if appropriate. A multilayer network typically includes three types of layers: an input layer which is a representation of the input data, one or more hidden layers which actually modify the data and an output layer which converts the activations from the hidden layers to an output (see figure 2.16). A layer in which each neuron is connected to each neuron of the previous layer is referred to as *fully connected* [93] whereas layers in which only few neurons are connected to the previous layer is called *sparsely connected* [94].

2.4.2 Training a Neural Network

During the learning process, the network parameters are updated to achieve the intended task. The aim is to adjust the parameter values of weights w_i associated with the connections between layers and biases in order to minimize an error function measuring the discrepancy between the desired output and the output of the model. However, in a multilayer network, it is difficult to solve the optimization analytically. Therefore, iterative numerical procedures can be used to find a sufficiently good solution. A possible approach is to use an iterative gradient descent optimization, previously described in more detail in section 2.3.5. This approach is able to deal with various activation and error functions. To start, the weights of the neural network are initialized to small random values and the input is propagated forward through the network. The received network output is then compared to the desired output using a loss function. To update the network parameters, the gradient of this loss function is calculated with respect to the weight parameters of the network. This is done using the chain rule of derivatives to propagate the error back through the network. Thus, the calculation of the gradient proceeds backwards through the network, with the gradient of the final layer of weights being calculated first and the gradient of the first layer of weights being calculated last. This is called *back propagation* [95]. The neuron weights are finally updated by subtracting a proportion of the gradient of the weights from the weights. This proportion is called the *learning rate*. By repeating the procedure, the network model is able to learn the given task.

Since this process corresponds to an iterative optimization procedure, different methods beside

gradient descent optimization can be employed to update the weights, similar or corresponding to those presented in section 2.3.5. These can include methods to estimate an appropriate step size for the weight change, since the network might take too long to converge if the step size is too small, or never converges and starts to oscillate if the step size is too large. Other approaches focus on weight decay, which scales down all weights after every iteration to avoid weights growing improportionally large [96, 97].

A common problem when training an artificial network is the so-called *overfitting*. In this case, the network is perfectly trained for a given task, but is not able to recognize general input. This can be monitored by dividing the training data into three subsets. These subsets correspond to one set of training data, which is actually used to train the network, and two subsets, called the validation and test data, which are held-out during training and only used to evaluate the model performance. This is done, since testing the network model skill on already known training data would result is a biased performance score. The difference between the validation data and the test data is that the validation data is used to estimate the model performance during training, however the resulting error is not taken into account for updating the network parameters, whereas the test set is used to evaluate the final model. When the error increases in the validation set, this is an indicator for overfitting.

Since the data sets are often too large to be processed as a whole, especially in image processing, the data is provided to the network in *batches*. The batch size then defines the number of samples which are worked through before the neural network model updates its parameters. Another important parameter during training of neural network is the number of *epochs* whereas an epoch defines the number of times the algorithm will work through the entire data set which is used for training.

Although the concept of multi-layer neural networks exist since 1980s, the training of large neural networks only became possible in the last few years. This is linked to different reasons, such as the lack of sufficiently large training data and powerful computers. The technological advances of graphics processing units (GPUs) allowed to realize the training of complex deep neural networks. Moreover, the development of novel activation functions such as ReLU or novel layer types that reasonably decrease the amount of data further advanced the realization of various DL methods and thus opened a way for the development of novel image processing methods.

2.4.3 Convolutional Neural Networks

The main challenge in processing image data is that even modestly sized images contain an enormous amount of information. An average 2D medical image contains for example $256 \times 256 = 65536$ pixel and for 3D images, this number additionally increases by two orders of magnitude. Thus, the amount of required weights to process such a 2D image in a fully connected neural network would also amount to 65536, if the information of each pixel is separately processed. This would result in a computationally demanding and slow training process. Moreover, image processing tasks often require information from surrounding pixels. An approach to integrate the information of neighboring pixel in classical image processing is the use of well-known convolutions such as e.g. Gaussian operators, Laplacian operators or gradient operators like a Sobel filter.

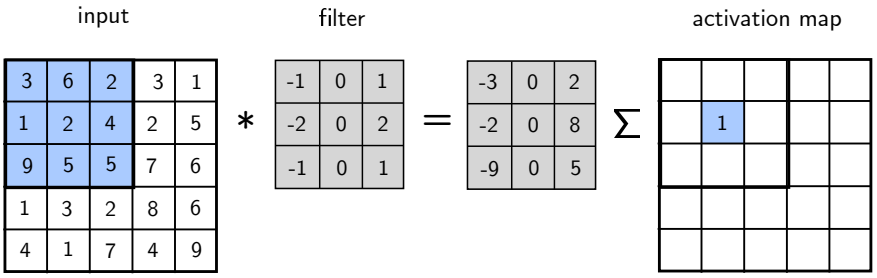


Figure 2.17: Principle of a convolution filter. The activation map is determined by sliding a filter, whose entries are learning during training, over the input image. The matrix values correspond to the dot product of the filter and the sub-image that is covered by it (here shown in blue).

To handle this vast amount of data in DL, a similar approach is used by introducing the concept of convolutional neural networks [98, 99]. The main difference to fully connected neural networks is that convolutional neural networks rely on information processing based on a convolutional filtering of the input of neighboring neurons. Figure 2.17 displays an example of a convolutional filter whereas the size of the receptive field, which corresponds to the region in the input space that a particular feature of the neural network is affected by, is defined by the size of the filter kernel. The discrete convolution of a 2D image f with a filter kernel g is defined as

$$h(x,y) = f(x,y) * g(x,y) = \sum_n \sum_m f(n,m) g(x-n, y-m) \ . \tag{2.30}$$

Basically, the dot product of the filter g and a sub-image of f with the same size as the filter centered at the coordinates (x,y) results in the intensity value of the filtered image h at pixel (x,y). By shifting the filter over the entire image f an output matrix h is created. In the context of convolutional neural networks h is referred to as *feature map*. The procedure requires a special treatment of the pixels at the image borders, since the convolution filter exceeds the image boundaries in this case. The image can either be padded by adding pixels with e.g. zero intensity or the intensity of the closest edge pixel, or the size of the resulting feature maps decreases slightly with every convolution since the calculation neglects border pixel in the input which are not covered entirely by the convolution kernel.

By applying such a set of convolutional filters, a convolutional layer in a neural network can be created whose filter kernel values are trained as conventional neuron parameters. The height and width of the output feature map depends on the dimensions of the input to the layer map whereas its depth depends on the number of applied filters. Since these filters are used for the entire image, the number of required parameters is drastically reduced compared to a fully connected layer in which each input channel is treated separately. So a fully connected layer uses one weight for each input pixel, whereas a convolutional layer only requires one weight per element of the filter kernel. The neurons of the convolutional layer share the same parameters and thus ensure translation invariance.

By connecting convolutional layers with other types of layers, a convolutional neural network (CNN) is formed. Typically other types of layers which are connected include *pooling layers*. Pooling layers are often used to decrease the size of the activation map created by the previous layer to decrease

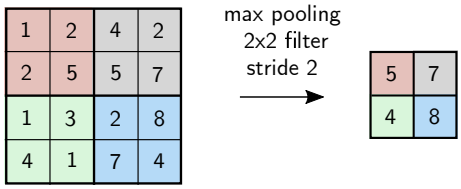


Figure 2.18: Maxpooling results in a down-sampling of the input image.

computation time. This is done, because deep layers of the network often require less spatial information about exact feature locations, but more filter matrices to recognize high level patterns. This can be realized by implementing a pooling layer with a *stride* larger than 1 (otherwise the size of the output of the pooling layer is the same as the input size) to down-sample the size of the activation map. The stride controls the distance between the centers of the applied filter kernel. If the filter is applied for every pixel in a neighborhood, the stride equals 1, if the filter is only applied to every n-th pixel the stride equals n. In addition, the introduction of a pooling layer increases the translational invariance of the network. A typical pooling filter is a *max pooling layer* which outputs the maximum value within a rectangular neighborhood of the activation map as shown in Figure 2.18. Another method to reduce the size of the activation map is to employ a *stride* larger than 1 during the convolutional operation. For certain applications, it is necessary to increase the size of the activation maps again, which is done by using an *upsampling layer*. Here, the size is increased e.g. by surrounding each value with zeros or a copy of this value.

Another way to regularize the learning process is the employment of *dropout layers*. The main idea of a dropout layer is to randomly disable input units by setting the outputs of hidden neurons to zero at a pre-set probability after each training iteration. Due to these dropouts, the network samples a different architecture every iteration which forces the neurons to learn more robust features.

A special type of CNNs are Fully Convolutional Networks (FCNs). They are built only from locally connected layers such as convolution, pooling and upsampling layers, but not fully connected layers. This results in a reduction of the number of parameters and computation time and allows for a computation which is independent from the original input size since all connections are local.

2.4.4 Encoder-Decoder Architecture

A very popular network architecture for image processing tasks are *encoder-decoder networks* which, as the name implies, consist of two stages: an encoding and a decoding stage. The encoder is typically composed of a sequence of convolutional and pooling layers and is used to embed the information contained in an image in a compressed lower dimensional feature representation. After every pooling step, the following convolutional layer has an effectively increased receptive field taking into account a larger region in the original input image. This is important for the network to not only learn local features, but also global context for larger image regions. A decoder is build for the contrary purpose, thus, to decode the information contained in the lower dimensional feature space. In consists of a series of convolutional and upsampling layers and aims to reconstruct the compressed output from the encoder. The main idea to use these encoder-decoder networks is that by compressing the input to a small intermediate feature representation, the network has to

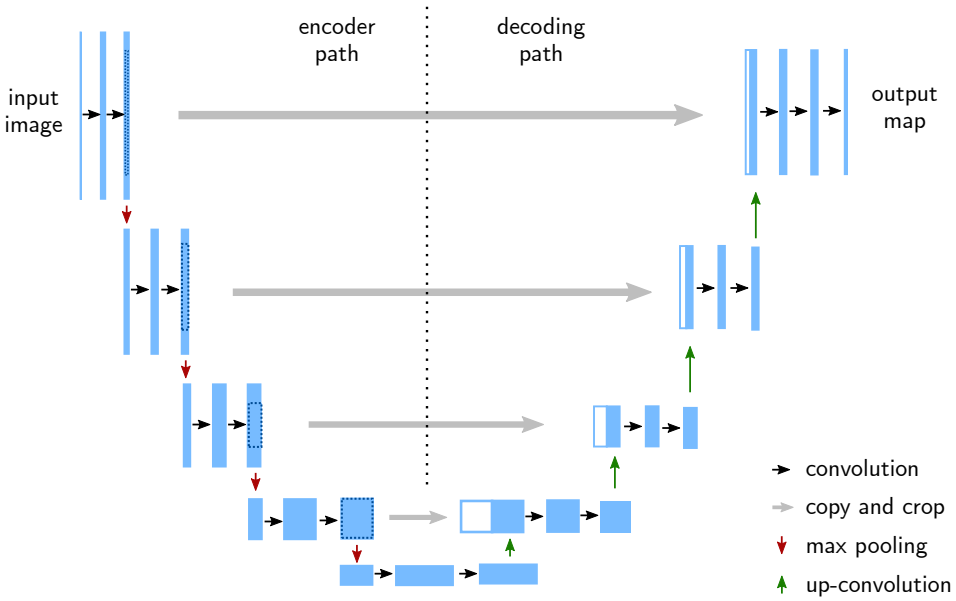


Figure 2.19: Encoder-decoder architecture based on the U-Net puplished by Ronneberger *et al.* [100].

learn a fitting representation of the input data to ensure that the decoder is able to reconstruct it correctly. These encoder-decoder architectures have proven to be more effective with small data sets as fully connected convolutional networks which are designed to learn little bits of information over a vast number of examples.

The most popular encoding-decoding network architecture is the *U-Net* proposed by Ronneberger *et al.* [100]. Its introduction has made a huge impact in many computer vision applications, especially image segmentation, outperforming conventional methods in various fields [101]. In addition to the standard encoder-decoder path, the U-Net includes "'short-cuts"', so-called skip connections, which are typically implemented as channel concatenations or additions between encoder and decoder at the same depth as shown in Figure 2.19. In this way, the decoder receives additional information and is able to regain spatial context and to include finer details of the encoder stage as well. This architecture helped to achieve impressive results in image segmentation [100] and has been successfully applied for many image-to-image translation tasks. New approaches also rely on this architecture to learn image registration tasks as will be explained in chapter 3.3.

In summary, the recent developments using DL in computer vision and especially medical image processing opened up an active field of research which yields new approaches for different image processing tasks. Moreover, the field of DL-based methods still offers a multitude of open questions including the investigation of novel network architectures or an effective characterization the impact of certain training parameters which is essential for the understanding of learning process.

3

State of the Art

The work presented in this thesis contributes to three different research areas in the field of medical image registration:

1. Image registration evaluation,
2. Novel feature-based similarity metrics for image registration and
3. Deep Learning in medical image registration.

Although image registration has been a challenging task in computer vision for many years, all of these topics are active fields of research and especially DL-driven registration methods are currently quickly developing and changing. Therefore, the current state of the art of these three research areas will be presented in this chapter, including a literature review of the most relevant approaches to our research.

3.1 Image Registration Evaluation

Despite the fact that there have been significant advancements and developments of registration methods themselves, techniques of registration evaluation received little attention.

Until today, there exists no standardized evaluation methodology to compare and benchmark existing registration methods. The main reason for this is the diversity of registration categories as shown in Section 2.3.1 which makes it difficult to define general criteria to evaluate the performance of a registration algorithm. The importance of certain criteria highly depends on the intended application scenario, e.g. real-time applications focus on time efficiency of a registration method whereas other applications require the highest possible registration accuracy in terms of image alignment.

Some of the characteristics which are considered relevant to evaluate an image registration algorithm and are therefore most often discussed in publications are:

- **Image alignment:**
  Since image registration aims to optimally align two or more images, the accuracy and precision of this alignment represent the most important characteristics. The accuracy refers to

the error/misalignment obtained for ideal input data and the precision refers to variations of this error.

- **Capture range:**
  A similar requirement is represented by the demand for a high capture range of the registration technique. The capture range generally defines the magnitude of image displacements, that can be corrected by the registration method. Hence, a large capture range ensures that the algorithm is able to find a geometric transformation to realign the input images, although the initial displacement between both images is very large. This is especially significant for rigid registration methods which aim to find global geometric transformations.

- **Plausibility of transformation:**
  This is an especially important point for deformable image registration, since the algorithm should output a physically plausible and smooth displacement field without folding or tearing. Moreover, the transformation determined by the registration method should maintain inverse consistency. This means that the estimated transformation should not be affected by the choice which image is used as reference and which as source image, but should maintain symmetry such that $T_{A\to B} = T^{-1}_{B\to A}$. If a third image is included, the transform should also maintain transitivity when registering image $A \to C$ so that the overall transformation corresponds to the joint transformations $T_{A\to C} = T_{A\to B} \circ T_{B\to C}$.

- **Robustness and Stability:**
  A registration method should be robust in more than one sense. For once, the results of the registration algorithm should not be significantly affected by perturbations in the image data, such as different degrees of noise or small artefacts. In addition, the algorithm should be robust in terms of consistency/stability, yielding a deterministic outcome when repeatedly executed on the same data.

- **(Computational) Requirements:**
  This refers to the required preprocessing and demand for IT resources of the algorithm. Both of these requirements are directly related to the time efficiency of the algorithm. Advanced complex registration methods or algorithms which rely on manual or automatic preprocessing can be time-consuming whereas time-constraints highly depend on the intended application.

Another very significant cause which makes the development of a standardized registration evaluation difficult is the fact that there rarely exists perfectly aligned ground truth data which can be used for the validation of the image alignment. Although monomodal registration methods can theoretically be evaluated using the same input data as reference and geometrically transformed as source image, there exists no ground truth data for the registration of two independent monomodal scans. This is even more significant for multimodal registration methods, for which the input data is acquired with different devices. As a result, the lack of a gold standard evaluation in terms of alignment accuracy is a major limitation in the field of non-rigid registration and multimodal registration. This is especially crucial for medical image registration, since a validation and evaluation of registration results represents an important step before introducing a registration method to

clinical routine. Thus, the absence of an evaluation gold standard impedes the use of image registration in different clinical workflows. Nevertheless, several surrogate measures have been developed aiming to estimate registration accuracy in terms of image alignment which will be explained in the following.

A very simple evaluation methodology to estimate registration accuracy is to apply a known deformation to an image, and then attempt to recover it by registration [102, 103, 104]. By comparing the known deformation field to the one estimated by the registration approach by using e.g. a distance measure, it is possible to evaluate the registration process. This evaluation approach conveniently quantifies actual registration errors (and thus accuracy), but is limited by its inability to quantify the accuracy of registrations between two independent images. Moreover, the generation of synthetic displacement fields often ignores physical properties of the organ tissues and therefore results in non-realistic deformations.

Most commonly employed alignment measures are based on the global overlap of different structures in the images and require image preprocessing in form of the generation of segmentation masks [105, 106, 107]. One of these measures is the Dice Similarity Coefficient (DSC) [108]. DSC is an overlap measure of labeled image regions E (for example given by binary segmentation masks) between the transformed source image $I_s(E)$ and the reference image $I_r(E)$ and is given as

$$DSC = \frac{2|I_s(E) \cap I_r(E)|}{|I_s(E)| + |I_r(E)|} \ . \tag{3.1}$$

A high DSC score (typically $\geq 90\,\%$) indicates a good overlap between the labeled regions after image registration. An alternative to the DSC is the Jaccard Coefficient (JC) [109] which is defined as

$$JC = \frac{|I_s(E) \cap I_r(E)|}{|I_s(E) \cup I_r(E)|} \ . \tag{3.2}$$

Both of these overlap measures can be related through

$$DSC = \frac{2JC}{JC + 1} \tag{3.3}$$

and can be calculated as single value for the overlap of two segmentation masks including several structures or they can be calculated for the mask overlap of each structure separately.

Another accuracy measure that is also based on the estimation of global overlap is given by the Hausdorff Distance (HD) [110]. On contrast to the DSC, HD does not rely on labeled image regions but on finite sets of points $A = \{a_1, ..., a_p\}$ and $B = \{b_1, ..., b_p\}$ in the source and target image. The HD is then defined as

$$HD(A, B) = \max(h(A, B), h(B, A)) \tag{3.4}$$

with

$$h(A, B) = \max_{a \in A}(\min_{b \in B} \|a - b\|) \ . \tag{3.5}$$

The HD measure is often used to estimate the distance between a contour in the deformed source image and the target image. A low HD value corresponds to a low distance between these points and therefore a good registration accuracy.

However, is has to be noted that these approaches highly depend on the quality of the employed segmentation masks. Moreover, using only one of these accuracy measures can be misleading, since a high DSC can indicate a good region overlap whereas the HD measure for the same registration can be very low due to an inconvenient contour registration. Therefore, a combination of different accuracy measures can be very useful. A selection of these surrogate measures were part of the software package Nonrigid Image Registration Evaluation Program (NIREP) [111] which included methods to estimate ROI overlap, intensity variations, inverse consistency and the transitivity error, but which is no longer available.

A major drawback of overlap measures and contour alignment measures is the fact, that they do not evaluate the registration accuracy within the labeled region. Using the example of non-rigid liver registration, such an accuracy measure relying on a segmented liver mask and its contours may indicate that the liver region in both images is perfectly aligned, but does not quantify the correct alignment of vessel branches or other structures inside the liver. The use of similarity metrics as presented in Section 2.3.4 is also not appropriate to measure alignment accuracy, since these methods are often biased to registration algorithms which employ the same or a similar similarity metric in their cost function.

Rohlfing [112] impressively demonstrated that commonly used accuracy measures such as tissue label overlap scores or image similarity measures are not sufficient to evaluate deformable registration performance. In his paper, he showed that that even combinations of these measures are unable to distinguish between accurate and inaccurate registrations. He did this by implementing a registration method which purposely yields inaccurate registration results but which could nevertheless achieve high accuracy scores using these image similarity and tissue overlap measures. Thus, these methods do not provide valid evidence for accurate image registration. According to Rohlfing, the standard for reporting nonrigid registration errors should ideally be reported as actual registration errors measured at a large number of densely distributed landmarks (i.e. distinct identifiable anatomical locations).

Point-based accuracy measures rely on point correspondences in the images that are registered. If these points are represented by anatomical landmarks or fiducial markers, two types of error are important: The Fiducial Localization Error (FLE) and the Fiducial Registration Error (FRE). The FLE corresponds to the distance of a localized point from the actual fiducial position in an image before any alignment is done [113]. Thus, it gives an estimation of the identification accuracy of a fiducial. The FRE then defines the Euclidean distance between the center of a fiducal in the source image which is affected by the FLE and the center of its corresponding fiducial in the reference image.

If the point-based accuracy is not limited to fiducials but includes the Euclidean distance between a point in an image and its true position, the error is referred to as Target Registration Error (TRE) [114]. Mathematically expressed, the TRE is given as the root mean square distance of homologous targets $t_{s,i}$ and $t_{r,i}$ with $i = 1, 2, .., N$ in the source and reference image, respectively

$$TRE = \frac{1}{N}\sqrt{\sum_{i=1}^{N} \left(T(t_{s,i}) - t_{r,i}\right)^2} \ , \tag{3.6}$$

with the geometric transformation T of the source image. Although the definition of FRE and TRE are quite similar, Fitzpatrick *et al.* [115, 113] showed that both error types are in fact uncorrelated and can not be used interchangeably. The points which are used for the calculation of the TRE are typically chosen in a target location of some treatment and are therefore often more clinically-relevant than the FRE [116]. However, due to some inconsistencies in terminology, it is not uncommon that fiducial-based registration errors are referred to as TRE in literature.

Point- or fiducial-based registration evaluation methods are suitable for monomodal as well as multimodal image data depending on the type of landmarks. One way to obtain the point sets for registration evaluation is by manually choosing salient points in the data sets [117, 118, 119]. But this is highly time-consuming and the results are biased by the accuracy of the annotations. Some projects published their annotated data to help others develop registration evaluation methods, such as Vandemeulebroucke *et al.* [120] whose data set consists of a single 4D lung CT data set which is composed of 10 3D images with 41 landmarks each or Castillo *et al.* [117, 121, 122] who also published 10 4D lung CT data sets with 300 landmarks per respiratory phase within the frame of the Deformable Image Registration Laboratory (DIRLab). The Retrospective Image Registration Evaluation (RIRE) project used skull-implanted fiducials in patients to generate a gold standard transformation for the evaluation of multimodal rigid registration techniques and published their data in a open-access database [123]. However implanted markers are highly invasive and in practice, the manual determination of convenient target points or non-invasive fiducials is challenging. Therefore, Murphy *et al.* [124] developed a semi-automatic process to generate ground truth for point-based registration evaluation of thoracic CT scans and also published 30 pairs of thoracic CT scans within the frame of the Evaluation of Methods for Pulmonary Image Registration 2010 (EMPIRE10) project [125]. In the EMPIRE10 challenge, registration algorithms are evaluated based on lung boundary alignment, fissure alignment, correspondence of manually annotated point pairs, and the presence of singularities in the deformation field.

Most of the presented approaches to medical image registration evaluation focus on the monomodal registration of head or thoracic scans. Two of the rare approaches which focus on the registration of abdominal scans are presented by Xu *et al.* [126] and Lee *et al.* [127]. In these papers, six and five different registration methods were compared for the registration of abdominal CT scans, respectively. However, the results on both publications are based on tissue overlapping accuracy, using the DSC and HD as surrogate measures which are arguable accuracy estimates as discussed previously.

Evaluation of Image Similarity Measures

Instead of evaluating the performance of an entire end-to-end registration method, it sometimes is necessary to benchmark the performance of an isolated component of a registration algorithm to investigate its influence on the overall performance of the method. An especially important component in most registration methods is the employed similarity measure. Traditionally, the behavior of an image similarity measure is investigated either indirectly by studying the quality of the final registration outcome or by sampling the parametric space of the metric obtained for the application of transformations relative to a ground truth. The latter results in the acquisition of a

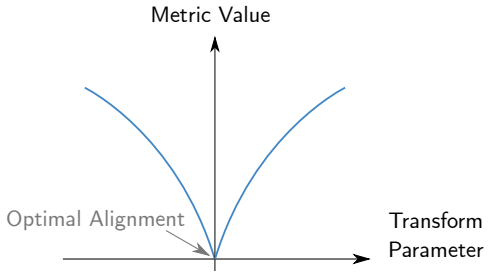


Figure 3.1: An ideal parametric cone is characterized by a smooth surface without local extrema and a clear extremum indicating the optimal image alignment.

parametric cone that generally describes the landscape of the similarity measure in its parametric space. In an ideal case, this landscape should be characterized by a smooth surface without local extrema and a clearly identifiable extremum at the position of optimal image alignment as shown in Figure 3.1.

However, the performance evaluation based on the parametric sampling is naturally limited to a fraction of the parametric space. This lead to the development of more advanced similarity evaluation techniques for rigid [128] and non-rigid registration [129]. Although both methods still rely on the sampling of the parametric space, they extend the traditional sampling by examining specific properties of the parametric space, including the distinctiveness of the optimum, the capture range and the number of local extrema. However, current approaches to investigate the performance of image similarity measures neglect the influence of differences in the image size of the input data or the position of optimal image alignment on the similarity estimation.

Contributions of this work:
Up to now, there exists no freely available evaluation methodology for the registration of multimodal abdominal scans, which relies on point-based accuracy measures. Therefore, such an evaluation methodology was developed within the frame of this thesis, focusing on the accuracy evaluation for the registration of 3D CT, MRI and CBCT scans of the liver. The proposed method is based on the evaluation methodology for 2D-3D multimodal registration of vertebral bodies presented by van de Kraats et al. [130]. However, the approach was expanded to 3D-3D registration, including a modified preprocessing to set the focus on the registration of the liver. The developed evaluation methodology will be presented in Section 4.2. To further investigate the influence of the employed image similarity measure on the registration outcome, an additional evaluation step is presented. This evaluation step extends the sampling of the parametric space of a similarity measure by distinguishing between different types of input data and thus, enables an appropriate comparison of similarity measures for specific input images. This similarity metric evaluation is presented in Section 4.2.4.

3.2 Novel Multimodal Similarity Metrics

As presented in Section 2.3.4, the definition of a multimodal similarity metric is highly challenging, since there exits no functional relation between the intensity mapping of corresponding anatomies

in different image modalities due to their differences in physical acquisition principles. Up to date, similarity measures based on Mutual Information (MI) as defined in Eq.2.20 and its derivatives represent the standard metrics for multimodal image registration tasks [5, 69, 73, 70]. Nevertheless, these approaches suffer from several disadvantages. In general, MI is known to be non-convex and typically results in local extrema which hinder the parameter optimization during image registration. This may result in incorrect image alignments or an increased computation time. Moreover, MI is estimated based on the joint grey value distribution in both images. This requires the use of approximations of the intensity distribution that are complex to compute and may involve sensitive smoothing parameters. But the main drawback of MI-based similarity measures for non-rigid registration tasks is the fact that MI-based metrics are intrinsically estimated on the entire image and thus rely on global instead of local information. As a consequence, MI-based methods have been very successful for rigid image alignment, but their application for non-rigid registration tasks is more demanding. Thus, novel approaches to define multimodal similarity metrics have been published in the last years.

Metrics based on Scalar Images

Mellor *et al.* [131] e.g. proposed the concept of phase mutual information. Instead of modeling an intensity mapping between two images, this approach models a relationship between local image phase and uses the extracted features as basis for a similarity measure. Other approaches focus on the information gained by the calculation of image gradients [132, 133, 134]. One approach which is relevant to some experiments performed within the frame of this thesis is the *Normalized Gradient Field* (NGF) similarity measure published by Haber *et al.* [135]. In this approach, the normalized image gradients are defined as

$$\tilde{\nabla} f = \frac{\nabla f}{\sqrt{\|\nabla f\|^2 + \epsilon^2}} = \frac{\nabla f}{\|\nabla_\epsilon f\|}, \tag{3.7}$$

whereas ϵ corresponds to an edge parameter that controls the impact of image gradients and should be chosen in the range of edges of interest. The NGF loss can then be estimated as

$$L_{NGF} = \int_\Omega 1 - \left(\tilde{\nabla} f(\mathbf{x} - \mathbf{u})^T \cdot \tilde{\nabla} g(\mathbf{x})\right)^2 d\mathbf{x} \ . \tag{3.8}$$

Here, $f(\mathbf{x})$ and $g(\mathbf{x})$ correspond to the intensity values at position $\mathbf{x}$ in the source and target image, respectively, whereas Ω defines the entire image domain and $\mathbf{u}$ the displacement.

In general, the employment of image gradients for similarity estimation represents a reasonable approach, since derivatives result from sudden intensity changes in the image which stem from the image structure and are thus related to organ/tissue boundaries. Wachinger *et al.* [136] also rely on structural information by proposing the definition of structural image representations which enable the use of L1 or L2 distance as similarity measure.

An advantage for the optimization of the cost function during registrations relying on these approaches is the fact that the image representations can be minimized using point-based differences.

However, for complex multimodal input data, scalar image representations may not be discriminative enough which lead to the development of similarity measures which are based on more discriminative feature descriptors.

Metrics based on Discriminative Feature Descriptors

A popular feature representation is the Self-Similarity Descriptor [137]. As the name suggests, the descriptor relies on the concept of self-similarity which defines the similarity within an image and can be estimated patch-wise using simple measures such as SSD. The patch-wise computation results in a local descriptor which can be extracted and matched across images [138]. The general assumption is that even though there exits no linear relation between the intensity distribution of anatomical features in two images, the intensity distribution within a local neighborhood is reliable and similar in different images of the same object. Consequently, the Self-Similarity Descriptor incorporates local information and is independent of the particular intensity distribution in the images making it suitable for multimodal deformable image registration.

This simple concept has been exploited in many different areas of image analysis and has been expanded to more advanced and robust versions which can be used as basis for an image similarity measure. One approach is to use a Local Self-Similarity Descriptor in combination with MI [139, 140]. Another derivative of the Self-Similarity Descriptor is e.g. the so-called Modality Independent Neighborhood Feature Descriptor (MIND) presented by Heinrich *et al.* [141]. The MIND encodes the local configuration of the Gaussian-weighted patch-distance in a neighborhood of the encoded voxel and is characterized by an advanced spatial sampling of the neighborhood, making the descriptor robust to noise and illumination changes. By comparing the descriptor of different images using simple monomodal distance measures, a similarity between these images can be estimated and used for image registration. Although this approach has been validated successfully for CT-MRI and MRI-Ultrasound registration tasks, the MIND is rotationally invariant which is a limitation for the registration of strongly rotated images. Although self-similarity seems to be the most commonly used concept to extract a discriminative feature descriptor as basis for medical image registration, there exists another approach which relies on a binary image descriptor referred to as discriminative Local Derivative Pattern (dLDP) [142]. dLDP is calculated as a binary string for each voxel according to the pattern of intensity derivatives in its neighborhood and evaluated using the Hamming distance, instead of conventional L1 or L2 norms.

In this work, a novel feature-based similarity metric is propoposed which relies on *Histograms of Oriented Gradients* (HOG) features. Originally, HOG were proposed as basis for a feature extraction algorithm applied for human detection in 2D photos or videos [143] and are now used for various computer vision applications [144, 145, 146]. The feature detection algorithm is based on computing gradient orientation histograms and results in a global feature descriptor (or feature vector) which can successfully describe the contents of the image regarding its gradient information.

The original version of the HOG descriptor is designed for 2D images. The first step to determine the HOG descriptor of a 2D image is to calculate the gradients in x- and y-direction using a derivation kernel, e.g. a Sobel filter. The magnitude g and the direction θ of the gradient in each

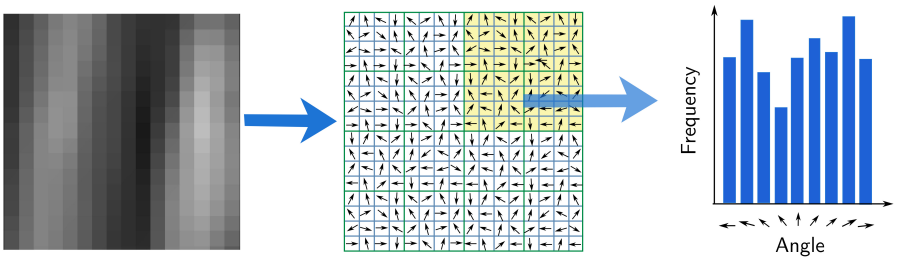


Figure 3.2: Generation of a Histogram of Oriented Gradients. After the gradient computation, the image is subdivided into cells, shown in red, that are grouped to overlapping blocks such as the one exemplary shown in yellow for normalization.

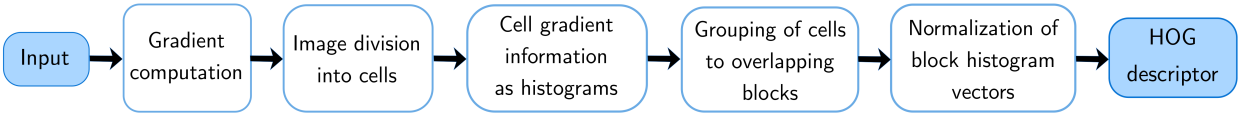


Figure 3.3: Extraction flow of a HOG feature descriptor.

pixel can then be calculated using the following formulas:

$$g = \sqrt{g_x^2 + g_y^2} \tag{3.9}$$

$$\theta = \arctan \frac{g_y}{g_x} \tag{3.10}$$

The gradient orientation θ is generally defined as angles between 0° and 180° instead of 0° to 360°. It has been empirically shown that these *unsigned* gradients result in an increased robustness for detection tasks based on HOG descriptors [143]. Next, the image is divided into equally sized subregions called cells and the cells are organized in overlapping blocks as exemplary shown in Figure 3.2. For every cell, a gradient histogram is calculated in which each histogram bin covers a certain angular orientation range. e.g. in the original paper, a single bin corresponds to an angular range of 20°. The transfer of the gradient information into a histogram results in a quantization and thus, a compression of the information.

The histogram entries correspond to the weighted gradient magnitude whereas the magnitude is split between two neighboring bins. Under the assumption that the bin size is 20° and the gradient orientation corresponds to 85°, $\frac{1}{4}$ of the gradient magnitude is added to the bin centered at 70° and $\frac{3}{4}$ of the gradient magnitude is added to the bin centered at 90°. It is important to note that histogram entries for angles close to 180° may be partially transferred to the first bin, since 0° and 180° are considered to be equivalent. The aim of weighting the histogram entries is to take gradients into consideration which are right on the boundary of two bins. Otherwise a small shift of a gradient with a large magnitude would have a strong impact on the outcome of the histogram.

By organizing the cells in overlapping blocks and concatenating their histogram vectors, it is possible to apply a normalization of the resulting block histogram vector. This normalization is based on a larger region as the normalization of individual cell histograms. This leads to an increased robustness of the feature descriptor extraction against local shadowing or intensity variations is shown by Dalal *et al.* [143]. The final HOG descriptor for the entire image is then defined by

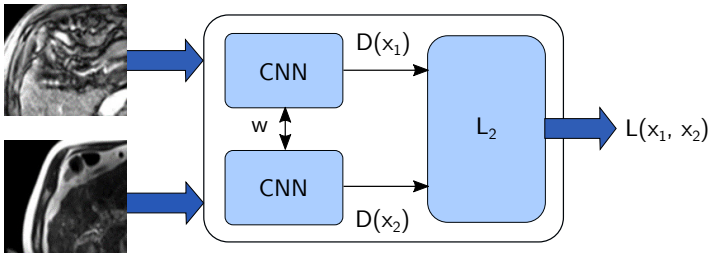


Figure 3.4: Scheme of a siamese network.

concatenating all normalized block histogram vectors to a single vector. The whole extraction flow is summarized in Figure 3.3.

In terms of image registration applications, only few approaches relying on HOG features exist. Abraham *et al.* [147] implemented an approach for point-based image registration by estimating and matching HOG feature descriptors of significant keypoints to generate a transformation field, whereas Zhou *et al.* [148] applied HOG for a rotational pre-alignment of 2D images before performing a registration.

Metrics based on Learned Features using Neural Networks

An alternative to the classical extraction of a feature descriptor comes with the rise of DL in the field of computer vision. Novel methods based on neural networks enable the learning of distinct feature descriptors which bear the potential to be used as basis for a multimodal image similarity measure. Examples of these approaches are given by the Fully Convolutional Self-Similarity (FCSS) descriptor presented by Kim *et al.* [149] and the Binary Robust Independent Elementary Features (BRIEF) descriptor [150]. However, the extraction of features which can be used to calculate a similarity value is not the only application of DL in terms of the development of novel image similarity measures. Novel methods also rely on the use of neural networks to directly estimate similarity between images without additional measures. This is called *Deep Metric Learning*.

More recent approaches to multimodal Deep Metric Learning employ a discriminative concept in which the learning of image similarity is treated as a classification problem that aims to discriminate between aligned and misaligned images of different modalities. Cheng *et al.* [151] propose a fully-connected CNN pretrained with an autoencoder to directly learn image similarity of multimodal 2D images as binary classification. The input is preprocessed using a trained denoising autoencoder network to learn appropriate feature representations which are more similar across modalities than the original images. This method is similar to the one presented by Simonovsky *et al.* [152] who scaled this approach to 3D input data and employs a training from scratch. The two input patches are concatenated and fed to a feed-forward CNN whose output corresponds to a scalar classification score.

Instead of concatenating the input images, approaches relying on *Siamese networks*, process the input data separately. A siamese net is a network which consists of two branches that share exactly the same architecture and the same set of weights to learn an optimal feature representation of two images before comparing these representations via loss function. Deep Metric Learning based on

siamese networks belongs to the group of supervised learning methods that are trained by using labeled data. For a siamese network, this data is represented by a pair of images with a binary label: the pair is labeled with 0 if the images are 'similar' and 1 if they are 'dissimilar'. The training of these networks often relies on a *contrastive loss* which is a combination of an Euclidean loss and a hinge loss, given as

$$L_c(x_1, x_1, y, m) = \frac{1}{2} y (D_W)^2 + \frac{1}{2}(1-y) \left\{\max(0, m - D_W)\right\}^2 \ , \tag{3.11}$$

whereas D_W represents the Euclidean distance between the outputs of each network branch

$$D_W(x_1, x_2) = \sqrt{(G_W(x_1) - G_W(x_2))^2} \ . \tag{3.12}$$

Here, x_1 and x_2 correspond to the input images, y to the assigned binary label and m represents a margin value larger than 0. Dissimilar pairs with a distance larger than this value will not be considered for the calculation of the loss. This ensures that the network is optimized learning the nuances to distinguish between images that are very similar, but do not display the same object. The contrastive loss forces a small distance between images that are labeled as similar and favors large distances between dissimilar images pairs. Thus, instead of learning to classify its network inputs, the neural network learns to differentiate between two inputs based on their similarity. The network is then trained by updating the weights of each network branch independently and then average the resulting weights. By feeding two input images in a trained siamese network the value for the contrastive loss for these two images is determined and gives a indication on the similarity of both images, thus serving as metric value. Although most siamese networks employ an encoder-decoder architecture for each branch, the exact architecture can be adjusted to the task at hand. Siamese networks are increasingly used for Deep Metric Learning [153, 154] and recent methods expand this techniques to Triplet networks [155, 156].

The approaches to learn a similarity measure using neural networks yield promising results, however, the main challenge of Deep Metric Learning in the context of medical image registration is the lack of sufficiently registered multimodal ground truth data.

Contributions of this work:
The contributions of this work to the research field of novel similarity metrics are twofold:
1) A novel similarity metric is proposed which is based on the use of *Histograms of oriented Gradients* descriptors to define a novel measure for multimodal image similarity. Recent approaches that incorporate gradient information as basis for a similarity measure typically rely on a global computation Therefore, a patch-based approach to integrate local information for registration purposes is proposed. The proposed similarity measure relies on the gradient orientation and magnitude and is implemented and evaluated for 3D-3D registration.
2) A siamese neural network architecture was implemented and trained in a medical context learning a multimodal similarity measure using T1- and T2-weighted MRI brain scans as well as synthetically generated MRI, CT and CBCT abdominal scans. The theoretical background as well as implementation details for both of these novel similarity measures are presented in Section 4.3.1 and chapter 4.3.2, respectively.

3.3 End-to-End Registration Learning using Neural Networks

Deep Learning benefits registration not only in terms of the evolution of novel similarity metrics, but also in terms of the development of novel methods for end-to-end registration learning. In the last few years, various new methods using neural networks in the field of medical image registration have been proposed and a summary of current approaches is given by Haskins *et al.* [157].

End-to-end registration learning represents a subcategory of Deep Learning methods in image registration. These methods aim to learn the spatial mapping of one image to register it with another. The output of these networks therefore either corresponds to the warped source image, the complete geometric deformation field which is necessary to align the images or both. One of the main advantages to use neural networks for image processing is the fact that once the networks are trained, they allow for a very fast processing of the input images. This can be especially important for time critical applications including image registrations in an interventional scenario.

In general, end-to-end registration learning defines the registration as parametric function and optimizes its parameters given a set of training data. The registration of new data pairs is then computed by evaluating the function using the learned parameters.

Supervised Registration Learning

Most approaches proposing neural networks to learn such a parametric function for medical image registration rely on ground truth deformation fields for training their networks [158, 159, 160, 161]. This ground truth data is either obtained by simulating and applying deformation fields or by using classical registration methods. Although these supervised methods achieve impressive results, the requirement of additional information in form of warping fields limits the potential application of these algorithms, since it possibly restricts the deformations that can be learned.

Weakly and Unsupervised Registration Learning

Therefore, several weakly- and unsupervised methods have been published in the last years. These methods typically employ a CNN and a spatial transform function that warps one image to the other. Most of these approaches are driven by an image similarity measure which is used to compare the fixed input image with the transformed source image warped by the spatial transformer similar to conventional image registration. Examples for these types of networks are the so-called *DIRnet* presented by de Vos *et al.* [162] or the fully convolutional net presented by Li *et al.* [163]. Both networks lead to promising results for deformable registration tasks, but their demonstration is limited to 2D slices or 3D sub-regions of the original images as well as the correction of small deformations. A general disadvantage of similarity-driven methods is the fact that they inherit shortcomings of the employed similarity measure.

As for conventional image registration, feature-based methods represent an alternative to intensity-driven methods. These approaches aim to estimate the deformation fields from higher-level correspondence information in anatomical labels using an end-to-end trained CNN. Such networks that

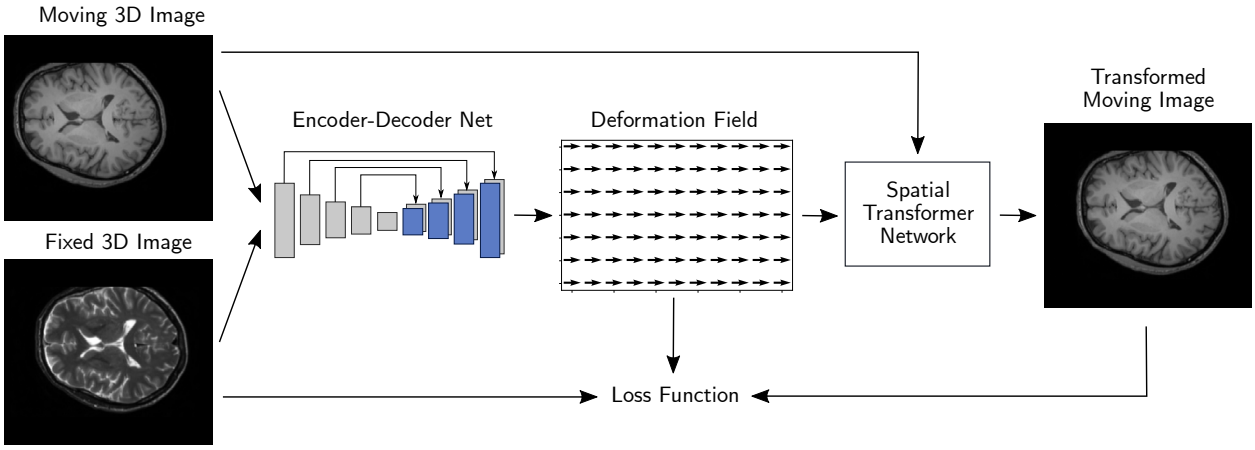


Figure 3.5: Overview of the VoxelMorph network.

predict a displacement field to align multiple labeled corresponding structures for image registration were presented by Wu *et al.* [164] based on a stacked autoencoder architecture and Yipeng *et al.* [165] relying on a weakly supervised CNN. As additional alternative to intensity-driven methods, approaches that rely on segmentation-based loss functions have been proposed. An example of such a method is presented and successfully applied for the multomodal registration of T2-weighted MRI to 3D ultrasound data by *et al.* [166, 167].

The *VoxelMorph* Network

An unsupervised learning model for deformable image registration which plays a particular role for the work presented in this thesis is the *VoxelMorph* network by Balakrishnan *et al.* [6, 168, 169, 170]. Once again, the parametric function is modeled based on a CNN and an additional spatial transform layer that processes both input images and outputs a mapping of all voxels from one image to the other while imposing smoothness constraints on the estimated registration field. In contrast to previously presented methods, this network is applicable for entire 3D volumes as well as large deformations. Moreover, the network does not necessarily require supervised labels for training and guarantees a diffeomorphic registration which preserves image topology. An overview of the method is shown in Figure 3.5.

The *VoxelMorph* network generally relies on an encoder-decoder architecture with skip connections, similar to a U-Net. The input to the network is represented by the source and target image which are concatenated to a 2-channel 3D image. The encoding and decoding are realized by applying 3D convolutions with a kernel size of $3 \times 3 \times 3$ followed by Leaky ReLU activation functions. For the encoding branch, the spatial image dimensions are reduced in half after every convolution due to a stride factor of 2. As for the conventional U-Net, this is intended to capture hierarchical features in order to estimate image correspondence. For the decoding branch, it is alternated between upsampling, convolutions followed by a Leaky ReLU activation and the concatenation of skip connections which propagate features learned during the encoding stage. The developers propose two variants of the network which differ in size at the end of the decoder stage and tradeoff between registration accuracy and computation time. The version VoxelMorph-1 uses one less layer at the final resolution and less channels over the last three layers [170].

The VoxelMorph network proposes two variants of registration learning: an unsupervised learning method using a loss function that only relies on the input volumes and the generated registration field, and an auxiliary supervised learning using a loss function that relies on anatomical segmentations. As for traditional image registration methods, the unsupervised version of the network is trained by minimizing a loss function that compares the warped source image to the reference image. The geometric manipulation of the source image is realized by using a so-called Spatial Transformer Network (STN) [171]. A STN performs the geometric transformation of an image as well as a linear interpolation of the intensity values, thus combining the image resampling and interpolator components of a traditional image registration algorithm.

The unsupervised registration learning relies on a loss function is constructed of two components: a loss L_{sim} that penalizes differences in image appearance and corresponds to a conventional image similarity metric, and L_{smooth} that enforces a smooth displacement field. L_{sim} is chosen as negative cross-correlation between the reference and source image. Since cross-correlation is a similarity metric mostly suitable for the comparison of images with similar intensity distributions, the registration learning using the unsupervised version of the original VoxelMorph network is limited to monomodal applications. L_{smooth} is introduced to encourage smoothness of the deformation field based on a diffusion regularizer on the spatial gradients of the transformation.

As an option, it is possible to rely on supervised learning by using an additional loss function based on the Dice coefficient to estimate the overlap of anatomical structures in the reference image and the warped source image during training. This requires annotated data sets with segmentation masks for both input images.

Initial experiments [6, 169] using the VoxelMorph network for the registration of T1-weighted brain MRI scans indicate that the network obtains a similar registration accuracy to state-of-the-art registration methods while reducing the computation time of the registration process by several orders of magnitude. A major advantage of the unsupervised VoxelMorph network over other registration learning methods is the fact that it does not require additional information such as ground truth registration fields or anatomical landmarks. Moreover, it is able to perform a registration of complete 3D image volumes instead of only smaller subvolumes.

In summary, novel methods for registration learning replace the costly optimization of traditional registration methods for each image pair by optimizing a global parametric function. This highly decreases the time required for an image registration, once the network is trained and thus, enables a multitude of new possibilities and applications. Since this is a rather new research field which developed quickly in the last few years, many new ideas and approaches are still to be expected.

Contributions of this work:
Within the frame of this work, the VoxelMorph network architecture has been modified to be used for affine and multimodal image registration tasks. Novel loss functions such as an unsupervised Siamese Deep Metric loss and a supervised Deformation field loss haven been implemented for training the network and evaluated for the registration of multimodal brain and abdominal scans.

4

Materials and Methods

In this chapter, the contributions of this thesis to the research field of medical image registration will be presented. The chapter starts with a general overview of the employed software libraries. Then the chapter is divided into three subsections each referring to contributions to the research areas in medical image registration including image registration evaluation, novel image similarity measures and end-to-end registration learning.

The first section focuses on evaluation methodologies. It starts with a presentation of the preprocessing of the image data used for the experiments that are performed within the frame of this thesis. The aim of this preprocessing is the generation of ground truth image data that can be used for the evaluation of image registration processes. Moreover, a strategy to evaluate the performance of similarity measures is presented as well as a novel evaluation methodology for the registration of multimodal abdominal scans. The developed evaluation framework is explained and used to compare and evaluate most commonly employed multimodal registration methods.

In the second section, two novel similarity metrics are presented. The first metric corresponds to a feature engineering approach and relies on the extraction of *Histograms of Oriented Gradients* as basis for a multimodal similarity measure. The second metric employs a *Siamese neural network* to learn image similarity as a classification task to distinguish between similar and dissimilar image patches. The presentation of both similarity measures includes details concerning their technical implementation as well as a presentation of the experiments to evaluate their performance on 3D-3D medical image data.

The third section is dedicated to study the performance of different variants of the *VoxelMorph* network for end-to-end registration learning. The original network is restricted to deformable monomodal image registration. Therefore, the network is modified to enable affine registration learning. Moreover, novel cost functions are implemented that are applicable on multimodal input data and thus, allow the application of the network for multimodal image registration. The performance of the original network as well as the modified versions of the VoxelMorph network are characterized for different application tasks and image data sets.

4.1 Toolkits and Hardware

The work presented in this thesis is mostly implemented in the programming language *Python*, except for the project presented in 5.3.1 which is implemented in *C++*. In order to simplify the process of algorithm implementation, the software is developed using a set of software libraries. These libraries and toolkits provide tools which facilitate basic image processing tasks or the implementation of neural networks and will be shortly presented in the following. Moreover, the employed IT-Infrastructure will be presented.

Software Libraries

The libraries which were employed for general image processing and analysis include:

- *ITK*:
  The Insight Segmentation and Registration Toolkit (ITK) [172] is an open-source, cross-platform application development framework for image registration and segmentation methods. It is developed in C++ and wrapped for Python, and relies on advanced templated programming which makes the code highly efficient and expandable to higher spatial image dimension and different pixel types. ITK contains all standard components used for image processing including e.g. different interpolators, transformations or metrics, and allows the development of individual modules which can be integrated in the intended application.

- *SimpleITK*:
  SimpleITK [173] is an offshoot of the ITK project which provides a simplified interface to ITK. This library is available in multiple programming languages, also including Python, and usable on all three major operating systems. SimpleITK only exposes the most commonly modified parameters settings of the ITK components, making it easy to use and thus, allows for a fast setup of image processing pipelines.

- *SimpleElastix*:
  SimpleElastix [174] is the python wrapper of for the *Elastix* tool [175]. Elastix is a software toolkit developed for medical image registration which provides a collection of highly optimized registration algorithms. It is based on ITK and its interface relies on a modular parameter file which can be executed via command line. SimpleElastix is the corresponding Python wrapper for Elastix which can be imported into Python as a module and enables a direct modification and execution of the parameter file.

- *MITK*:
  The *Medical Imaging Interaction Toolkit* (MITK) [176] represents open-source software for medical image informatics, image processing and interactive 3D visualization. MITK combines ITK and the Visualization Toolkit (VTK) [177] with an application framework and enables the development of interactive medical image processing software.

There exist different libraries to set up a functional neural network which are based on the same theoretical machine learning models, but differ in their approach on how to implement them. The work presented in this thesis is based on two libraries:

- *TensorFlow*:
  TensorFlow [178] is a Python library which was originally developed as part of the Google brain project and made open-source in 2015. It offers integrated tools such as TensorBoard, a visualization tool that automatically generates graphs of scalars from summary files, such as e.g. the loss function, the learning rate or model weights. Moreover, TensorFlow enables the distribution of workload on several GPUs, thus offering a computational advantage.

- *Keras*:
  Keras [179] is a neural networks application programming interface written in Python that runs on top of either TensorFlow, Theano, PlaidML or Microsoft Cognitive Toolkit which are all different software libraries for machine learning. Since it is designed modular and extensible, it allows for fast and easy prototyping of machine learning algorithms and minimises overhead.

Hardware

Due to the size of the input data and the large number of parameters optimized during image registration or contained in a DL algorithm, the use of an appropriate hardware that is able to handle this amount of data is essential to minimize computation time. For computationally complex tasks, GPUs are favorable over a CPU, since they represent highly parallel computing engines which enable a significant reduction in computation time compared to CPUs. The experiments presented in this work that are characterized by a high computational demand were therefore performed on a server equipped with two *Intel Xeon X5670* CPUs with 96 GB RAM as well as and three Nvidia GPUs: two *Titan Xp* GPUs (with 12 GB RAM) as well as one *GeForce GTX 1080 Ti* GPU (11 GB RAM).

4.2 Evaluation Methodology for Medical Image Registration

The overall goal of the work presented in this thesis is the development and optimization of image registration methods for multimodal abdominal scans. In general, the performance of a registration algorithm is highly dependent on the involved modalities, and therefore needs to be investigated for each modality combination. To optimize a registration method it is crucial to specify criteria which define the quality of a registration outcome. Since there exists no evaluation standard for image registration, an evaluation methodology for linear and nonlinear registration of multimodal abdominal scans has been developed.

The presented evaluation methodology includes two parts: The first part of the presented methodology refers to the evaluation of the entire image registration process. The outcome of each method

is evaluated in terms of registration accuracy. As discussed in Section 3, overlap-based evaluation scores are not suitable to sufficiently describe the performance of a registration method, thus, the presented methodology relies on a point-based accuracy estimation. The second part aims to discuss and implement a strategy to validate and compare different similarity measures. This is important since image similarity metrics represent the most crucial component of the registration method regarding multimodal image registration since there exists no linear relation between the intensity distributions in different modalities.

However, as described in previous chapters, the main limitation to evaluate existing and newly developed image registration methods as well as similarity measures is the lack of ground truth data in form of optimally registered image pairs. The only exception to this is the use of one single image simultaneously as reference and geometrically transformed as source image for the registration.

Therefore, the first step to implement a registration evaluation method consists of an advanced preprocessing of the employed image data to generate a sufficiently aligned ground truth. Parts of this evaluation and the required data preprocessing were published in the proceedings of the international conference *SPIE Medical Imaging 2019* [180].

4.2.1 Data Pre-processing and Generation of Ground Truth Data

To enable a performance evaluation of novel similarity measures and registration methods developed within the frame of this thesis, two approaches to artificially generate surrogate ground truth data were employed. With regard to the use-case of image registration for the context of liver interventions, the focus is set to the generation of abdominal ground truth data sets consisting of pre-interventional CT and MRI data as well as intra-interventional CBCT data. The first approach relies on classical image processing based on manually chosen landmarks whereas the second approach employs a neural network to generate synthetic abdominal MRI, CT and CBCT scans based on a digital phantom. Although multimodal ground truth data is scarcely available, it is possible to use intra-modal ground truth data as surrogate multimodal data for registration evaluation. This intra-modal image data is e.g. represented by T1- and T2 MRI scans of the same subject acquired during the same imaging session. Moreover, several experiments in this thesis aim to investigate the influence of the morphology depicted in the image data on the performance of similarity measures or entire registration processes. Therefore, a third ground truth data set is presented comprising T1 and T2 MRI of the brain.

Ground Truth Generation using a Point-based Pre-Registration

The first approach relies on the complete pre- and intrainterventional multimodal image data of patients that underwent a transarterial chemoembolization (TACE) in the liver. The preinterventional data is given as 3D T1- and T2-weighted MRI (T1 and T2) acquired on a 3 Tesla MAGNETOM Skyra MR Scanner[1]and a 3D CT acquired on a SOMATOM Force CT scanner[1]. A 3D CBCT acquired during the performance of the TACE on an Artis ZEEGO® system[1] serves as

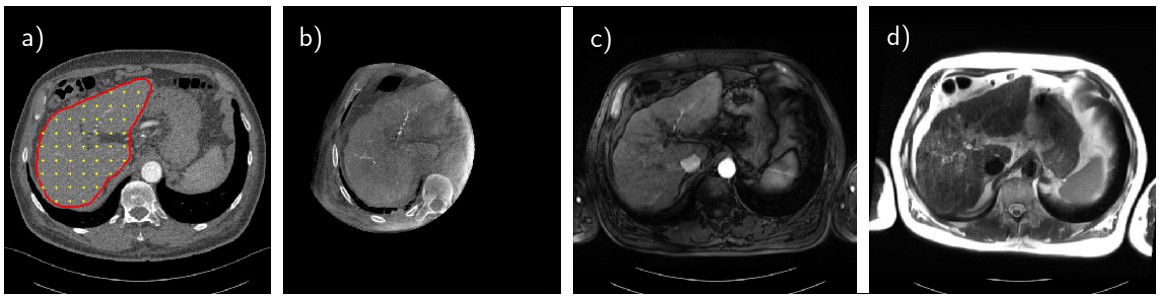


Figure 4.1: a) Axial slice of the 3D-CT data with mask (shown in red) and artificial landmarks (shown in yellow) for the estimation of the target registration error (see Eq.3.6); b) axial slice of the CBCT data; c) axial slice of the T1 MRI; d) axial slice of the T2 MRI.

Table 4.1: Details of the patient data used for the evaluation of linear registration methods.

Patient	Modality	Size [px]	Pixel Spacing [mm]
1	CT	$512 \times 512 \times 189$	$0.83 \times 0.83 \times 1.5$
	CBCT	$512 \times 512 \times 375$	$0.49 \times 0.49 \times 0.49$
	T1	$384 \times 312 \times 64$	$1.09 \times 1.09 \times 3$
	T2	$384 \times 308 \times 45$	$1.09 \times 1.09 \times 4.4$
2	CT	$512 \times 512 \times 188$	$0.98 \times 0.98 \times 1.5$
	CBCT	$512 \times 512 \times 375$	$0.49 \times 0.49 \times 0.49$
	T1	$320 \times 250 \times 72$	$1.31 \times 1.31 \times 3$
	T2	$448 \times 364 \times 41$	$0.92 \times 0.92 \times 6$
3	CT	$512 \times 512 \times 172$	$0.68 \times 0.68 \times 1.5$
	CBCT	$512 \times 512 \times 375$	$0.49 \times 0.49 \times 0.49$
	T1	$320 \times 240 \times 80$	$1.31 \times 1.31 \times 3$
	T2	$270 \times 148 \times 60$	$1.44 \times 1.44 \times 4.4$

intra-interventional data. Three exemplary data sets are chosen for this preprocessing. Axial slices of the corresponding data sets for one of the patients are exemplary shown in Figure 4.1 a) - d) and further details of the image data are presented in Table 4.1. Figure 4.1 b) displays the circular shaped field of view which is characteristic for CBCT data. In general, this limited field of view in addition to the increased noise level in CBCT data (as discussed in Section 2.2.1) can pose major problems for image registration processes involving this modality.

The preinterventional CT is chosen as the target data for the generation of the ground truth registration. This is useful for multimodal medical image registration, since the CT is characterized by a high contrast-to-noise ratio as well as a high spatial resolution. Moreover, the CT is not as prone to geometric distortions and artefacts as MRI which is often affected by inhomogenities of the magnetic field or susceptibility artefacts, as discussed in Section 2.2.2.

To attain an accurate ground truth for the registration evaluation, a rigid point-based registration of the T1, T2 and CBCT data to the CT data is performed. The landmarks used for this registration are manually extracted for each individual 3D volume by choosing 16 distinct points inside the liver, such as i.e. branches of supplying blood vessels. The point-based registration is performed using an iterative closest point algorithm [79] implemented in *MITK*. To get a first impression

[1]Siemens Healthcare, Forchheim, Germany

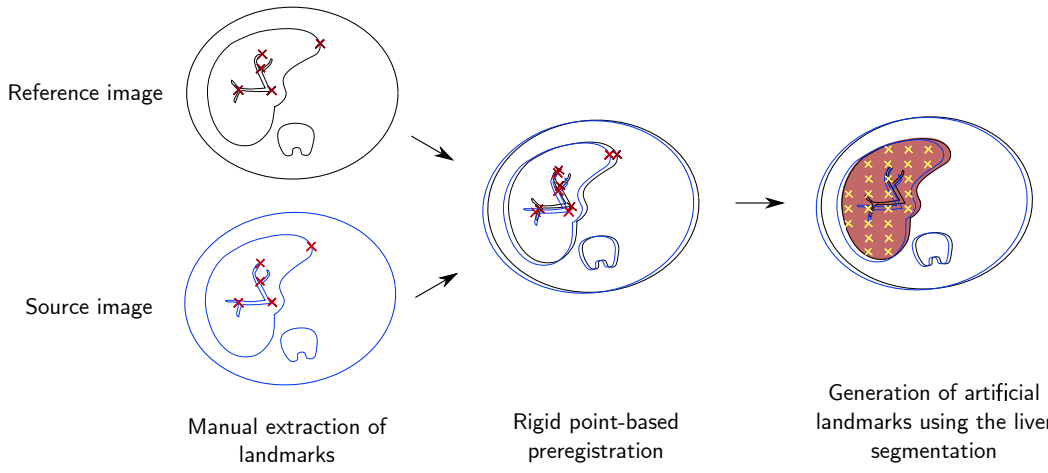


Figure 4.2: Visualization of the data preprocessing to generate a ground truth registration and the artificial landmarks used for the TRE calculation.

Table 4.2: FRE (in mm) of the manually chosen landmarks after performing a point-based rigid registration and after performing a point-based affine registration.

Modalities	Registration	Pat 1	Pat 2	Pat 3
T1 to CT	rigid	6.12	5.58	6.48
	affine	5.69	4.19	5.87
T2 to CT	rigid	5.66	7.72	6.36
	affine	5.48	7.65	5.23
CBCT to CT	rigid	6.43	5.02	4.16
	affine	5.25	4.78	3.36

of the magnitude of the image displacement after this preregistration, the FRE (see Section 3.1) before and after a rigid as well as an affine point-based preregistration is calculated for the manually extracted landmarks and listed in Table 4.2. These FRE values indicate the mean distance between corresponding landmarks after a rigid or affine preregistration and therefore, give an impression of the resulting accuracy and general quality of the preregistration.

In general, the manual extraction of distinct feature points is very time consuming and therefore not suitable for daily clinical routine. However, for the generation of ground truth data, the manual feature extraction and subsequent matching may result in a high accuracy which is mainly limited by the anatomical knowledge of the user and the spatial resolution of the images. Nevertheless, 16 landmarks per data set only result in a sparse displacement field between different modalities. Due to this limited number of spatial correspondences, the images are not aligned using a nonlinear transformation, since this potentially results in inappropriate interpolated voxel displacements between the manual landmarks caused by the lack of information in between these points. Since the sparsity of the manual landmarks could affect the possible evaluation accuracy, artificial landmarks are defined additionally.

As this thesis aims at the development and evaluation of new registration methods for the context of hepatic interventional procedures, the focus of the registration evaluation is set on the registration of the liver. Thus, the liver is manually segmented in the CT of each patient using *MITK*. Since this step is performed after the point-based pre-registration, the resulting segmentation mask is accurate

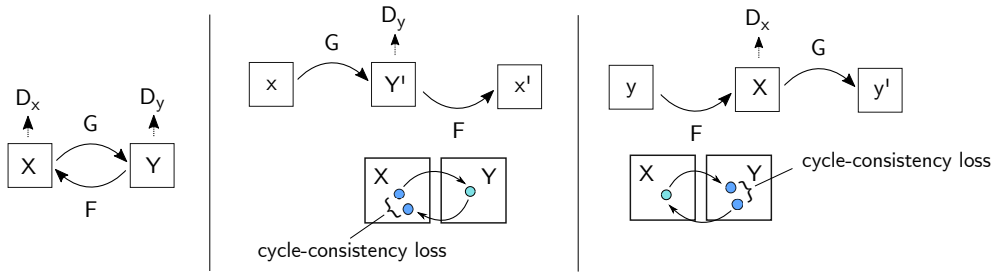


Figure 4.3: Illustration of the cycle loss employed in a CycleGAN. Image adapted from Zhu *et al.* [181].

for every modality of the corresponding patient. The mask is then used to distribute equally spaced points with a distance of 20 mm inside the liver serving as landmarks for the estimation of the mean target registration error (TRE) before and after registration. This results in about 200 artificial landmarks for each 3D volume. A visualization of the whole data preprocessing to generate the ground truth registration is shown in Figure 4.2. A slice through the segmentation mask for one of the patients and the resulting artificial landmarks are shown in Figure 4.1 a). An advantage of this procedure is the fact that it allows the estimation of dense displacement fields whereas the spacing between the artificial landmarks can be customized according to the registration task. However, the definition of artificial landmarks requires a sufficiently accurate pre-registration as presented here in form of the point-based ground truth registration.

This procedure results in three registered data sets consisting of T1, T2, CBCT and CT data of the same patient, whereas the latter serves as target volume, including 16 manual landmarks and about 200 artificial landmarks inside the liver as well as a segmentation mask for the liver.

Ground Truth Generation using a CycleGAN

A problem in using the patient data as ground truth data for deformable registration applications is the fact that the images are only rigidly aligned during the preprocessing. This leads to morphological deviations in the spaces between the manually chosen landmarks which result in an incorrect ground truth for nonlinear registration methods. Replacing the rigid-point based pre-registration with a nonlinear pre-registration is not reasonable, since this procedure leads to a bias in the registration evaluation favoring the registration method which is used for the pre-registration. Therefore, another approach to generate ground truth data is employed. This approach relies on the use of a special type of neural networks, so-called Cycle-Consistent Generative Adversarial Networks, to generate synthetic medical image data and was implemented by cooperation partners from the Department of Computer Assisted Clinical Medicine, Heidelberg University.

Cycle-Consistent Generative Adversarial Networks, mostly referred to as *CycleGANs*, gained much attention in the last years, since they allow the transfer of characteristics from one image to another while training with unpaired image data. They are a subtype of conditional Generative Adversarial Networks (conditional GANs) [182]. These networks are generally composed of two models, the generator and the discriminator. As the name suggests, the generator aims to generate new data from an input whose characteristics are similar to the characteristics of a second input. The role of the discriminator is to distinguish if an input is real or faked by the generator. The ultimate

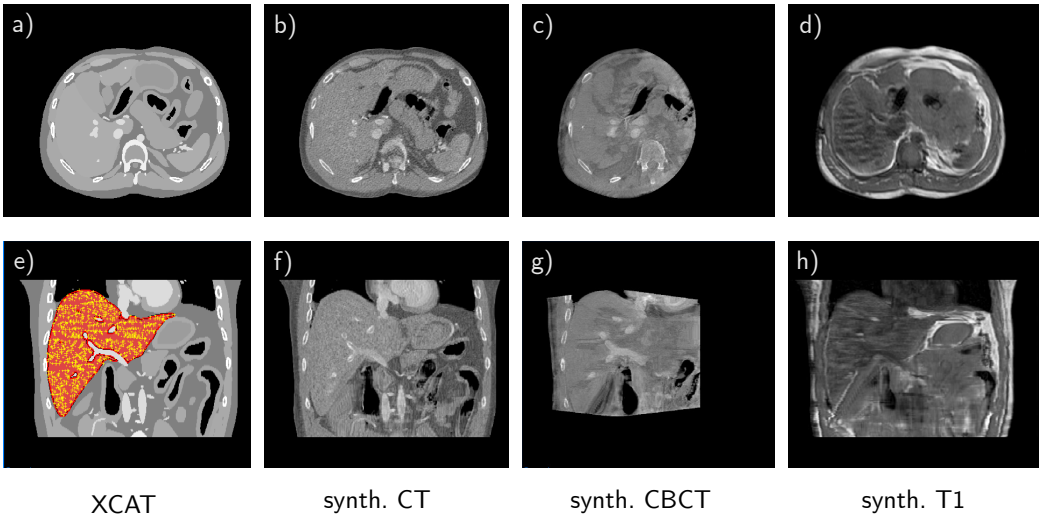


Figure 4.4: Axial and coronal slices extracted from the digital XCAT phantom a) + e) with mask (shown in red) and artifical landmarks (shown in yellow) as well as slices of the synthetically generated CT b)+f), CBCT c) + g) and T1 d) + h) data based on this phantom.

goal is to trick the discriminator such that fake images are classified as real images. Thus, these two models are trained in an adversarial fashion since they are trained to achieve opposing goals. One application field of GANs is the use for image-to-image translation tasks. Image-to-image translation generally describes the mapping from one specific image domain to another. A good example is the field of style transfer that refers e.g. to the reproduction of a photo in the style of a specific artist. However, this concept can also be transferred to medical image processing to generate synthetic data in a specified modality from a base modality.

The challenge of using conditional Generative Adversarial Networks for image-to-image translation is the requirement of pre-aligned, paired training data. To overcome this limitation Zhu *et al.* [181] introduced an additional cycle-consistency loss which enforces a similarity of an input image in domain X to the representation of the same image after mapping it to a domain Y and then back to domain X as shown in Figure 4.3. This network architecture is the so-called CycleGAN. Since it enables to use unpaired image data for training, this network architecture benefits various applications including medical image processing where pre-aligned data of different modalities is only sparsely available. First approaches to use CycleGANs for modality synthesis of medical data have been presented in [183, 184, 185].

The project partners from the Department of Computer-Assisted Clinical Medicine implemented a CycleGAN, additional details about their approach published in [186, 187]. In the context of medical image registration, this network is used for the generation of multimodal synthetic ground truth data. For the training of the network, we collected multimodal datasets of 18 patients that underwent liver biopsy, including preinterventionl T1 MRI, CT and intrainterventional CBCT abdominal scans. Compared to the patient data sets used for the ground truth generation described in the previous section, the data sets are composed of the same imaging modalities, however they differ in terms of the T1 MRI. The MRI sequence used as preinterventional data for the biospy differs from the one used as preinterventional data for the TACE leading to different image contrasts.

Table 4.3: Details on the XCAT image data as well as the synthetic modalities generated with the CycleGAN and used for the evaluation of nonlinear registration methods.

	Modality	Size [px]	Spacing [mm]
	XCAT phantom	$512 \times 512 \times 155$	$0.75 \times 0.75 \times 1.5$
Synthetic	CT	$512 \times 512 \times 155$	$0.75 \times 0.75 \times 1.5$
Data	CBCT	$512 \times 512 \times 386$	$0.49 \times 0.49 \times 0.49$
	T1	$512 \times 512 \times 155$	$0.75 \times 0.75 \times 1.5$

Instead of just training the network with patient data as two different input modalities, a digital antropomorphic phantom, the so-called XCAT phantom [188], is used as input. The digital phantom models the whole human body and allows for morphological variations by defining the organ size and position as well as the simulation of respiratory and cardiac motion. Since the focus of this thesis is set on the registration of abdominal image data, the XCAT phantom was used to model a set of multiple 3D scans of the abdomen by varying these parameters. These 3D volumes serve as input for the CycleGAN and patient data of a single modality is used as second input. Thus, the network is trained for each modality separately and learns a mapping from the XCAT phantom to the corresponding modality and vice versa. The trained network is then used to generate the ground truth data by feeding an XCAT data set in the network to generate a synthetic T1, CBCT and CT from this specific XCAT volume. Since the generation of all synthetic modalities rely on the same XCAT scan, they are intrinsically aligned and display very similar morphological features. Slices of the XCAT phantom and the synthetic modalities are exemplary shown in Figure 4.4. However, since the network is not perfectly optimized yet, slight morphological variances such as blurring artifacts are visible in the synthetic data sets. Nevertheless, by further optimizing this setup, this methods is very promising to help overcome the lack of ground truth data in medical image registration.

Another advantage of the XCAT phantom is the fact, that it allows to simulate the motion of a cardiac as wells as a respiratory cycle by producing nonlinearly transformed image data (including movement of the diaphragm, abdominal wall and the liver position). Moreover, this simulation yields the computation of corresponding voxel positions in the reference frame and transformed image frame of the phantom. This is very useful for the evaluation of nonlinear registration methods, since it provides a perfect ground truth for the calculation of displacement fields and allows the estimation of the TRE on a dense point grid. As additional information for image registration, a mask of the liver region can be easily generated by applying a threshold algorithm on the image data of the phantom. For the experiments performed in this thesis, the breathing cycle is simulated to extract the image data and displacement vectors for five positions on the sinusoidal breathing curve. Further information on the XCAT image data and the synthetic modalities used for these experiments is given in Table 4.3. The simulation outputs the fixed position of each voxel in the reference image as well as their new position in the transformed image leading up to several million point pairs. Since the computation of such a large point set is very time consuming and the focus of the registration evaluation is set on an optimal alignment of the liver, the data points are filtered to the points in the liver including only every hundredth data point. This still results in a dense point grid including about 20000 points inside the liver region. A coronal slice including the liver

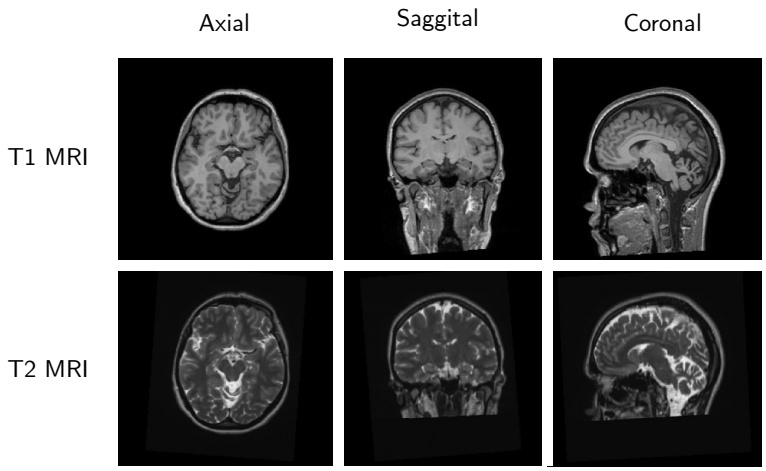


Figure 4.5: Slices through a pair of registered T1 and T2 MRI volumes of the IXI data set.

segmentation and landmarks is shown in Figure 4.4 e). As for the manually preprocessed data, these points serve as landmarks for the TRE calculation before and after the application of an image registration algorithm.

Intrinsically Registered Ground Truth Data

As third data set, the T1 and T2 MRI scans of the IXI data [189] are used. This data set is mostly used for all experiments in this thesis that require a large amount of image data, such as approaches that rely on DL, as well as for all expierments that aim to investigate the influence of the depicted patient morphology on the performance of similarity measures or registration approaches.

The IXI data set refers to a data collection of nearly 600 MR images of the brain from normal, healthy subjects including i.a. T1-, T2-and PD-weighted images. Since these scans are acquired on the same device, the images are intrinsically well-aligned and therefore suitable to serve as ground truth image data for evaluation purposes. Nevertheless, to correct for small image displacements between scans of the same subject, a rigid registration based on a mutual-information based similarity metric is performed using *SimpleITK* [173]. After the registration, all image volumes are resampled to a size of $256 \times 256 \times 160$ voxel with a spacing of $0.94 \times 0.94 \times 1.2\,\text{mm}$, so that all images display the same basic image characteristics. Although T1 and T2 scans are considered intra-modal data, they still yield different contrasts and display morphological structures differently as shown in Figure 4.5, thus, serving the purpose to be used as multimodal data sets.

In summary, three ground truth data sets are available, including:

1. **Real Abdominal GT Data:**
   Three real patient data sets, consisting of registered CBCT, CT and T1 MRI data, 16 manual landmarks and 200 artifical landmarks in the liver as well as a liver segmentation mask. This data set is used to evaluate approaches on real patient data. However, it is only suitable for approaches that do not rely on a large amount of data.

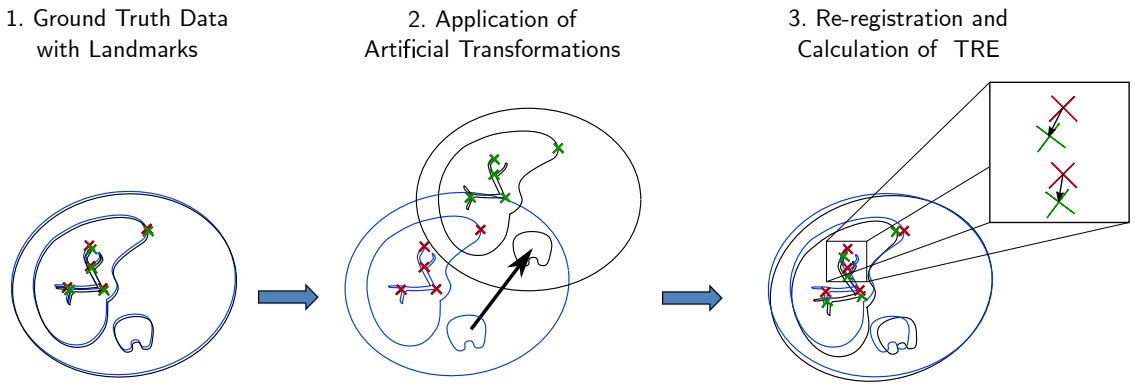


Figure 4.6: Overview of the methodology applied for rigid image registration evaluation. The steps 1. to 3. are repeated 450 times with different transform parameters. The TRE before and after the re-registration are determined and are used to estimate the capture range of the investigated registration method.

2. **Synthetic Abdominal GT Data:**
   Five synthetic abdominal data sets generated for five positions of the respiratory cycle, consisting of registered CBCT, CT and T1 MRI data, 20000 landmarks in the liver as well as a liver segmentation mask. This data sets is used for all experiments that require optimally registered multimodal images. Moreover, the data set serves for the extraction of a large amount of multimodal subvolumes that can be used to train neural networks.

3. **Real MRI Data of the Brain:**
   T1 and T2 MRI scans of the brain, including the data for 600 subjects. This data set is used to establish a feasibility study for all applications requiring a large amount of diverse registered data such as DL-based methods. Moreover, it serves as ground truth data for experiments which aim to investigate the influence of morphology on the registration outcome, by comparing the performance for abdominal and brain data.

4.2.2 Registration Evaluation Methodology for Multimodal Abdominal Data

In the first part of the evaluation methodology, the performance of an entire multimodal registration method is investigated. An evaluation for the multimodal registration of abdominal scans is implemented for linear and nonlinear registration methods, respectively. Both evaluation methods aim to analyze the performance of a complete registration algorithm in terms of registration accuracy and enable a comparison between different methods for the registration of the liver. Up to now, only few evaluation methods for the registration of abdominal scans are available which rely on segmentation-based measures such as the Dice Coefficient [108] or the Hausdorff distance [110]. However, overlap-based accuracy measures represent an arguable registration criterion. Therefore, both evaluation approaches for linear and nonlinear registration methods presented in this thesis rely on a point-based accuracy measure based on the estimation of the mean TRE using the landmarks defined in Section 4.2.1. The aim is to implement an evaluation methodology which enables an easy and fast comparison of different linear and nonlinear multimodal registration methods.

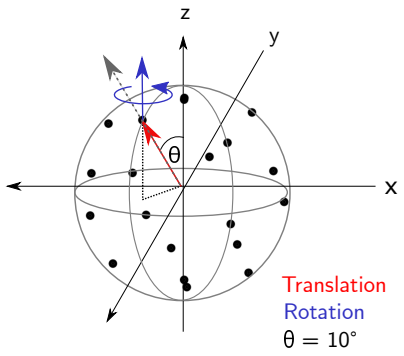


Figure 4.7: Transformations applied to the preregistered source images.

Linear Registration

Starting with the evaluation of linear image registration methods, the attainable accuracy of the different registration methods is estimated similar to the evaluation methodology presented by van de Kraats *et al.*[130]. The general idea is to start using an optimally aligned ground truth data image pair, apply a geometrical transformation to one of the images and correct the displacement by re-registering the reference and the transformed image. By calculating the mean TRE before and after registration, it is possible to evaluate the registration performance. Moreover, this procedure allows the estimation of the *capture range* of the registration method that is generally defined as the range of displacement positions from which the algorithm finds a sufficiently accurate transformation to realign the input images. An overview of the method is given in Figure 4.6.

Since the focus of this work is the optimization of registration methods for multimodal abdominal data, the data used for the evaluation experiments shown in this chapter corresponds to the pre- and intra-interventional image data of patients undergoing a TACE, further described in Section 4.2.1. For the evaluation of DL-based registration methods that require a higher amount of image data for training, the methodology is expanded for the use of the IXI data set as well as the synthetic abdominal data (see Section 4.2.1) applying the same preprocessing as described in the following. However, the experiments presented in the first part of this chapter focus on the application of the registration methods on real abdominal patient data to evaluate their performance in a accurate clinical context with regards to interventional procedures in the liver. After generating the ground truth registration of the other modalities to the CT target data for each patient, the source images are artificially transformed by applying rigid transformations using *SimpleITK* before re-registration.

These transformations include translation transformations in the range of $0 - 25\,\text{mm}$ in steps of $5\,\text{mm}$. The translation directions are set as vector from the image center to one of five points equally distributed on a spherical shell around the image center. These points are defined using the method of Saff *et al.*[190]. The transformation also includes rotations around the image center in a range of $-15\,^\circ$ to $15\,^\circ$ in steps of $7.5\,^\circ$. To take into consideration realistic variances of the patient positioning in the scanners, three rotation axis with a maximum inclination angle of $10\,^\circ$ relative to the axial direction of the patient are chosen. This multiplies to a total of $6 \times 5 \times 5 \times 3 = 450$ artificial transformations which are applied during the registration evaluation for each pair of image

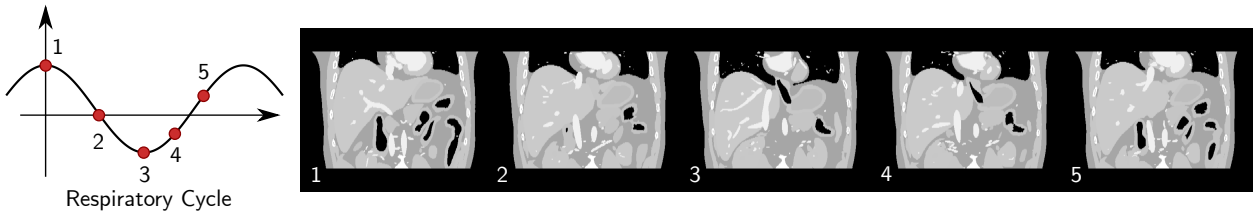


Figure 4.8: For the evaluation of nonlinear registration methods, the digital XCAT phantom was used to simulate a respiratory cycle and the image data at five positions was extracted and served as ground truth data for the evaluation.

modalities (CT/CBCT, CT/T1, CT/T2).

To avoid non-relevant black regions in the source image after the transformation, no image resampling is applied to obtain the transformed images. Since this could possibly influence the registration outcome as discussed in Section 4.2.4, only the image information, such as image origin and direction, is altered according to the applied geometrical transformation.

For the estimation of the capture range of each method, the mean TRE before and after a rigid registration of the target image and the geometrically transformed source image are calculated, using the ~ 200 artificial landmarks inside the liver.

Nonlinear Registration

The evaluation of nonlinear registration methods is based on the simulated data that is generated using the digital XCAT phantom described in Section 4.2.1. Compared to the patient data used for the evaluation of linear registration methods, this data is more suitable for the evaluation of nonlinear methods since it provides a higher alignment accuracy of the ground truth data.

The digital XCAT phantom is used to generate abdominal image data of a respiratory cycle at five different positions of the cycle, resulting in five abdominal XCAT volumes. The positions and a coronal slice of the corresponding XCAT volume are shown in Figure 4.8, whereas *position 1* is noted as reference frame. According to the preprocessing procedure based on a CycleGAN, each XCAT volume is used to generate a synthetic CT, CBCT and T1 MRI of the specific volume. Moreover, a list of voxel positions in the reference volume and their new position in the transformed image volumes is generated which serves as basis for the calculation of a mean TRE.

To estimate multimodal registration accuracy for nonlinear methods, each of the five synthetic CT serve as reference image in the registration process and are registered in a permutative manner to the synthetic CBCT and T1 MRI of the other respiratory cycle positions. Thus, the 3D images of positions 2, 3, 4, 5 are registered to the volume at position 1 in the first run, then the image volumes 1, 3, 4, 5 are registered to the volume at position 2 etc. This results in a total of 20 nonlinear registrations. As for the evaluation of linear registration methods, the mean TRE before and after registration are calculated to estimate the accuracy of each nonlinear registration method.

4.2.3 Experimental Setup

The image registration process for the evaluation of linear and nonlinear algorithms is performed using the open source software *SimpleElastix* [174]. *SimpleElastix* offers the possibility to use predefined so-called *parameter maps* for rigid, affine and deformable registrations. In addition, *SimpleElastix* enables an easy adjustment of individual parameter settings according to the intended registration task. This makes it a very suitable basis for the registration evaluation algorithm, since it allows a fast implementation of a various different registration methods. Nevertheless, the component of the presented evaluation methodology that performs the actual registration can be easily replaced by any custom registration method. This allows a comparison of commonly applied registration methods to very specific or newly developed registration approaches.

Linear Registration

For the evaluation presented in this work, the presets of *SimpleElastix* for a linear registration are used as basis for further optimization. The default parameter map generates a multiresolution registration with 4 levels using a linear interpolator, an adapative stochastic gradient descent optimizer and a three-dimensional Euler transform as geometric transformation. By default, Advanced Mattes Mutual Information [73] (AMMI) is employed as a similarity metric. To study different aspects influencing the outcome of the registration methods, the following parameters are further investigated for the multimodal registration of abdominal scans while keeping the other parameters at their default setting for every experiment:

- **Initialization**
  In most registration scenarios, a transform initialization is performed to prealign the images before starting the actual registration to reduce the geometric displacement which has to be corrected. This is often done by superposing the geometrical centers of the reference and source image. Based on the calculation of the mean TRE, this classic approach is compared to an initialization by aligning the geometric center of the target organ, in our case the liver. The geometric liver center has been chosen manually. These two types of superposition represent the first two initialization approaches investigated in this thesis. A third initialization is realized by aligning the liver center in the CT, serving as reference modality in these experiments, to the geometric center of the source image. This is especially interesting for the registration of CBCT to CT, since the liver is centered during the acquisition of CBCT scans acquired during a liver intervention. In contrast to the following experiments which rely on the mean TRE of the artificial landmarks, this comparison is based on the mean TRE of the manual landmarks, since the initialization is performed on the native image data and not the preregistered images.

- **Similarity metrics**
  Beside AMMI as default similarity metric, further metrics, namely Advanced Normalized Correlation [191] (ANC) and Normalized Mutual Information (NMI) [75], were applied.

- **Masks**
  The introduction of a binary mask as additional information for the registration methods was investigated. In general, masks can be used for the fixed and moving image to limit the evaluation of the similarity metric to the region of the intended target structure. This can be beneficial for the registration of organs highly affected by motion. In our case, the position and form of the liver varies due to the breathing motion. By using a mask, static significant structures such as the rib cage are not taken into account during registration. For the evaluation, two different masks were employed for the fixed image (CT), including the manual liver segmentation acquired during image preprocessing (Sec. 4.2.1), as well as a simple cuboid mask covering the liver region.

- **Number of resolution levels**
  In general, a higher number of resolution levels is expected to result in a higher capture range, but also requires a higher computation time which is an important factor considering interventional registration applications. To further study this tradeoff, the number of resolution levels was set to values between 1 and 4, whereas 4 layers correspond to the predefined parameter setting of the default parameter map for linear registrations offered by *SimpleElastix*.

- **Rigid vs. affine Registration**
  Due to the rigid preregistration, the liver is not perfectly registered and the metrics potentially optimize to a position which does not correspond to the position of the ground truth. To take this into consideration, the TRE of the manually chosen landmarks used for the point-based ground truth registration was estimated after applying a rigid registration and compared to the TRE obtained for an affine registration using the predefined parameter map offered by *SimpleElastix*. Similar to the parameter map for rigid registrations, this parameter map generates a multiresolution registration with 4 levels using a linear interpolator, an adapative stochastic gradient descent optimizer and an affine transform.

Nonlinear Registration

The deformable registration is performed using the default parameter map for B-spline-based image registration offered by *SimpleElastix*. The basics of this default parameter map are very similar to the default parameter map for rigid registrations. It generates a multiresolution registration with 4 resolution levels, employs an adaptive gradient descent optimizer and AMMI as similarity metric by default. The main difference is naturally represented by the usage of a B-spline interpolator as well as a B-spline transform as geometric transformation. For the registration evaluation experiments, different parameters of the nonlinear registration have been altered to further understand their impact on the registration of multimodal abdominal scans. The investigated parameters include:

- **Physical grid spacing**
  SimpleElastix enables the alteration of the grid density of the B-spline transform for each dimension individually by defining either a physical or voxel-based grid spacing. The basics of B-splines are explained in Section 2.3.3. In general, a higher number of grid points allows a higher flexibility of the transformation to model complex deformations. But although a

higher number of grid points results in a higher number of DOF, it is also equivalent with an increased number of transform parameters and therefore requires an increased computation time. If the number of grid points is chosen too high, the registration can result in unrealistic deformations of the source image, whereas a registration based on very few grid points is not able to correct complex deformations. Thus, the spacing of the grid points should ideally correspond to the size of expected deformations. To study its impact on the registration accuracy, the physical grid spacing was varied from 50 to 150 mm in steps of 20 mm.

- **Similarity metrics**
  The registration accuracy was estimated using AMMI, ANC as well as NMI as image similarity measure.

- **Masks**
  Similar to the evaluation of rigid registration methods, the influence of a fixed image mask on the registration accuracy of nonlinear methods was investigated using a binary mask of the liver as well as a cuboid mask that covers the liver region. These results were compared to the results obtained without a fixed image mask.

- **Number of resolution levels**
  The number of resolution levels was varied from 1 up to 4 resolution layers whereas the number of grid points for the B-spline transform remained the same for every resolution level. Since the physical spacing is increased by a factor of two for each level, this indirectly implies that the grid spacing is divided in half.

While most of the parameters were kept at their default setting for the performance of the experiments, the grid spacing of the B-spline transform in each dimension was adjusted to 110 mm for all experiments except the investigation of the grid spacing itself. This was done, since the default setting corresponds to a physical grid spacing of 8 mm which results in a very dense grid and therefore high computation times. Such a dense grid was not necessary for the respiratory deformations which were corrected via image registration.

4.2.4 Similarity Metric Evaluation

The experiments discussed in the previous chapter reveal the complexity of a registration method as well as the wide range of interdependent influences on the registration outcome. One of the most relevant parameters for the accuracy and robustness of a multimodal registration method is the employed similarity measure. Each similarity measure has different properties and is characterized by an individual sensitivity to the modality of the images, the image content, such as e.g. edges, interpolation, the size of the image overlap etc. This makes it desirable to estimate the quality of the similarity metric prior to registration. Thus, the second part of the evaluation methodology focuses on the quality assessment of similarity measures.

Evaluation Strategy
A common approach to investigate the behavior of an image similarity measure is the sampling

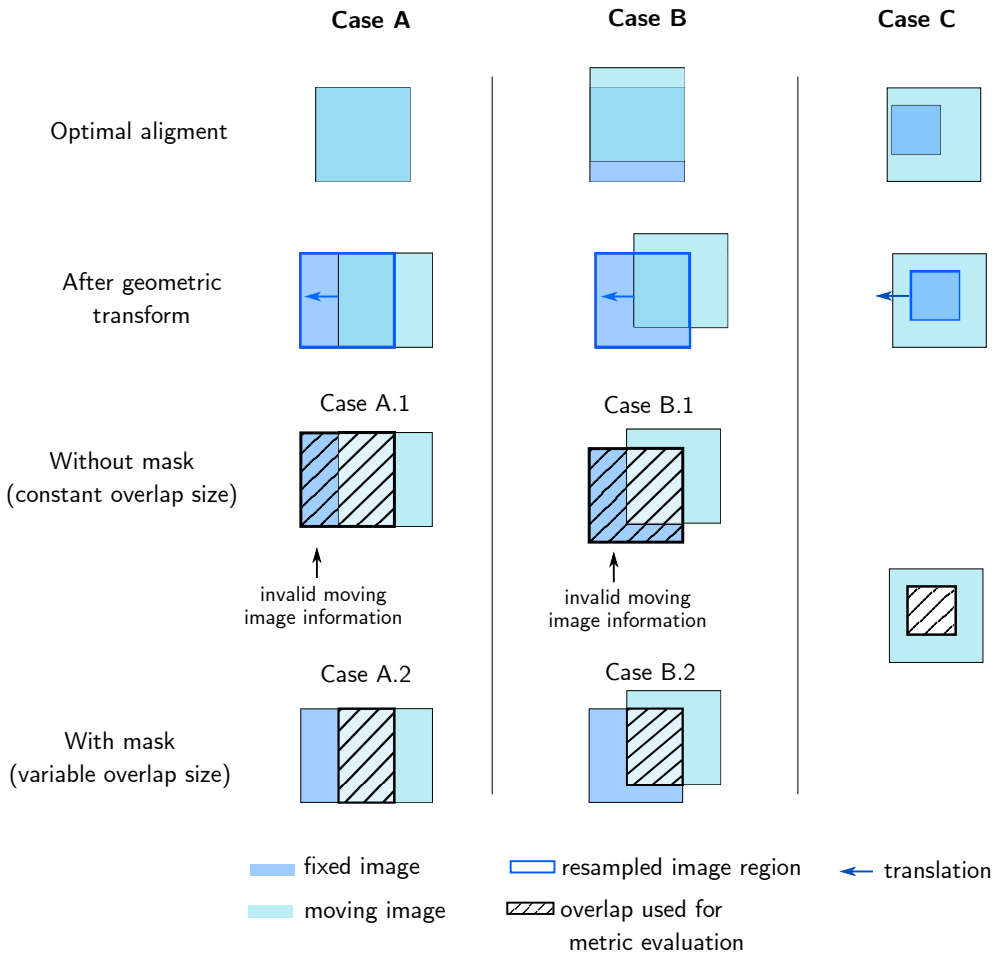


Figure 4.9: Visualization of the different cases considered for the evaluation of similarity measures depending on the optimal position of image alignment and the size of the overlapping image region in which the metric value is determined.

of the parametric space. This sampling for translation or rotation transformations is realized by determining the similarity metric values after shifting or rotating one image of an image pair relative to a gold standard position.

To evaluate the similarity metrics used and developed in this thesis, an evaluation algorithm has been implemented which facilitates the sampling of the parametric space by enabling shifts in a random spatial direction and image rotation around a specified rotation center and axis. The translation range used for the evaluation of similarity metrics corresponds to a shift of -30 to 30 mm. For most experiments, the metric values are estimated for a translation in x-, y- and z-direction. However, for specific cases, the parametric space is sampled in a random direction or an entire plane to further understand the behavior of the similarity metric. For the estimation of rotational transformation parameters, the rotation angle is varied between $-30°$ to $30°$ relative to the rotation axis corresponding to one of the three main axes or a specified rotation axis pointing in a random direction. The rotation center for all metric evaluations is set to the geometrical image center.

Different Cases

An important factor which is often neglected when evaluating a similarity measure is the dependence of the similarity metric on the position of optimal image alignment and the size of the overlapping image region. During a registration process, the source image is transformed interatively and the similarity value is estimated after every transformation step. Thus, the size of image overlap constantly changes during a registration process.

Case A:

Let us assume that the reference and source image have the same size and display the same object in the same image region. In this case, the position of optimal image alignment corresponds to the position of maximum image overlap (see Figure 4.9 *Case A*) and a simple registration could be performed by superposing the geometric centers of both images. This scenario is mostly relevant for deformable registration tasks, such as the correction of respiratory motion which require the correction of local organ displacements instead of a global organ alignment. A similarity metric would then be evaluated in the entire image region of both images, since they completely overlap.

However, if the images are not already aligned, there exist two different image regions which could be considered for the similarity estimation before alignment:

Case A.1:

The first region corresponds to the image region of the fixed image, taking into account an image region in the moving image which contains non-valid image values (visualized in Figure 4.9). In this case, the size of the image region used for the similarity evaluation is constant, however non-valid image regions are taken into account for the similarity estimation yielding a biased metric value.

Case A.2:

In a second scenario, the image region which can be used for metric evaluation corresponds to the overlapping region of both images (also visualized in Figure 4.9). In this case, only valid image information is taken into account for the similarity estimation between both images. However, the size of the image region that is considered is variable and depends on the current image alignment, thus having a direct impact on the resulting metric value. In real applications, this consideration of only valid image regions is realized by using binary image masks that mask out non-valid image regions. To minimize the influence of the overlap size on the estimated similarity value, a normalization using the number of pixels in the overlapping image region is usually required.

Case B.1 and Case B.2:

Another case is given for images, in which the same object is not located in the same image region and that require actual image registration to generate an overlap of the same structures in both images. In such a case, the position of optimal alignment does not correspond to the position in which the overlapping region of both images is maximized. This is exemplary shown in Figure 4.9 *Case B*. In addition to the problem of considering non-valid image information (in the following referred to as *Case B.1*) that can be overcome by using binary image masks (referred to as *Case B.2*), this offset could potentially influence the performance of the similarity measure, since some metrics tend to optimize for a maximal image overlap.

Table 4.4: Three different cases defined for the evaluation of a similarity measure depending on the optimal image alignment as well as on the size of the overlapping image region in which the similarity value is determined. A visualization of these cases is given in Figure 4.9.

Case	Mask	Image Overlap	Non-Valid Image Information
A.1	no	constant	yes
A.2	yes	variable	no
B.1	no	constant	yes
B.2	yes	variable	no
C	n.a.	constant	no

Case C:
Dropping the assumption that both images have the same size may result in a third scenario in which the image overlap remains constant during the sampling of the parametric space without the introduction of invalid image regions. This scenario is also illustrated in Figure 4.9 *Case C*. In such a (ideal) case, the similarity evaluation with and without a mask will make no difference, since the region in which the similarity metric is determined stays constant either way. In a clinical context, this scenario is important for the registration of images with significantly different sizes. An example for such a case is e.g. given by the registration of CBCT to CT scans, since CBCT generally covers a smaller field of view as CT.

In general, it is important to take into consideration the size of the image overlap during similarity metric evaluation. To study the behavior of similarity measures for all scenarios summarized and listed in Table 4.4, the implemented algorithm for metric evaluation enables the possibility to distinguish between these different cases, according to the optimal image alignment and the image overlap during the evaluation. Depending on the image registration task, the image data used as gold standard for this part of the evaluation is represented by one of the ground truth data collectives presented in Section 4.2.1.

This procedure allows to evaluate and even visualize the performance of a similarity measure on different imaging data in dependence of the specified transformation parameters. Moreover, the presented similarity evaluation method includes the investigation of different evaluation cases that enable a performance evaluation for very specific types of input images. In general, the acquisition of parametric cones gives a good impression of the efficiency of a similarity metric for a specific task. This makes this method an appropriate tool to facilitate the choice of an optimal similarity metric for a specific application as well as the development and optimization of novel similarity measures.

4.3 Novel Similarity Metrics

The results of the evaluation of different registration methods described in Section 4.2.2 show that the choice of similarity measure highly influences the achievable registration accuracy. Especially for multimodal image registration tasks, the choice of a suitable similarity measure is crucial. This chapter presents novel similarity measures and the strategy of their evaluation for multimodal

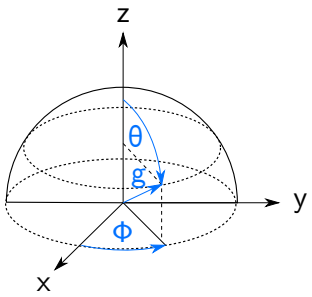


Figure 4.10: The gradient information in 3D is given as magnitude g and orientation defined by two angles θ and ϕ.

medical image registration.

4.3.1 Metric based on *Histograms of Oriented Gradients*

As first alternative to classic image similarity measures, a novel gradient-based 3D-3D-registration method is proposed. The proposed method relies on the use of *Histograms of Oriented Gradients* (HOG) as basis of a similarity estimation for the registration of pre- to intrainterventional image data. The metric is based on the patch-wise computation of gradient orientations and magnitudes and therefore does not rely on pixel-to-pixel but region-to-region correspondence of local distributions of gradient intensities and orientations. Hence, the metric is suspected to be less affected by local changes of absolute intensity values and to yield stable registration results. The implementation of the metric, the experimental setup and the results of this project have been presented as oral talk at the international conference *SPIE Medical Imaging 2017* and have been published in the corresponding conference proceedings [192].

Concept of a Similarity Measure based on 3D-HOG Descriptors

Although there exist various approaches using 2D-HOG descriptors in different application fields, there exists no application of *3D*-HOG descriptors for medical image processing tasks, including image registration.

The main motivation to use a pure version of a HOG descriptor directly as a similarity measure for gradient-based image registration is the important role image gradients (or edges) play in conveying image content. Moreover, the estimation of histograms of oriented gradients provides a high sensitivity towards translation, rotation and scaling transformations making it suitable as basis for a similarity metric. Moreover, gradient information generally relies on sudden intensity changes which are characteristic for organ margins in medical images. Thus, the proposed metric aims on the alignment of object outlines instead of the establishment of a grey value relation between the images, it offers the potential to be used on mono-as well as multimodal image data. Though initially proposed for 2D applications, the feature detection using HOG has to be extended to three dimensions for the intended task of 3D-3D-medical image registration. Therefore, a 3D-HOG descriptor was implemented and used as basis of a similarity measure.

The extension to 3D is realized by referring to the spherical coordinate system to describe gradient

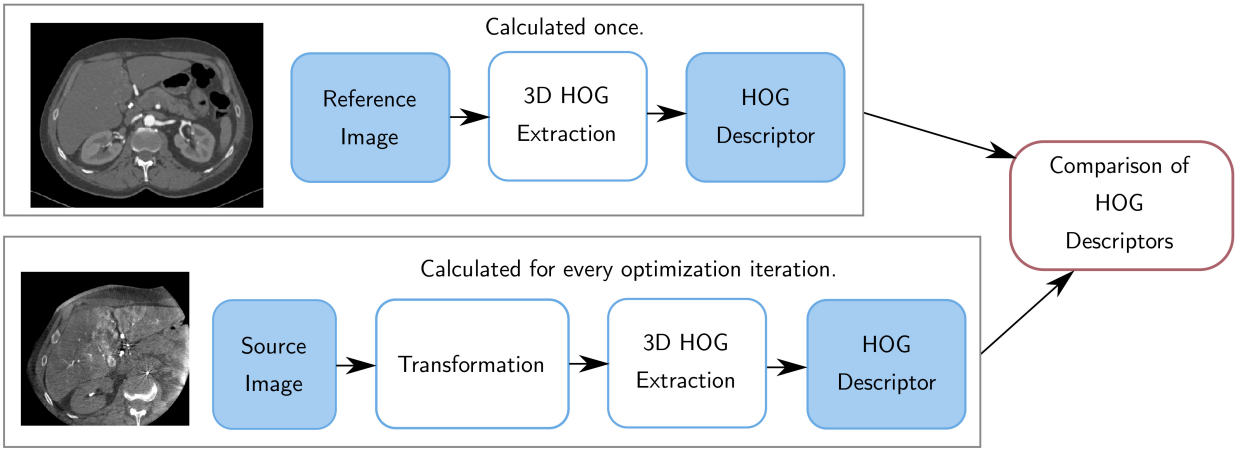


Figure 4.11: Application of *Histograms of Oriented Gradients* (HOG) as basis for a similarity measure. The HOG descriptor for the reference image is only calculated once whereas the HOG descriptor for the source image is calculated for every optimization iteration of the registration process.

orientations as shown in Figure 4.10. Similar to the 2D approach, gradient filters are applied for each spatial direction and the gradient orientation, given by two angles θ and ϕ, and magnitude g are estimated using the following formulas:

$$g = \sqrt{g_x^2 + g_y^2 + g_z^2} \tag{4.1}$$

$$\theta = \arctan \frac{g_y}{g_x} \tag{4.2}$$

$$\phi = \arccos \frac{g_z}{\sqrt{g_x^2 + g_y^2 + g_z^2}} \tag{4.3}$$

Both orientation angles cover different angular ranges due to the nature of 3D spherical coordinates. Whereas θ ranges from $0°$ to $180°$, ϕ covers an angular range from $0°$ to $360°$. However, to increase the robustness of the resulting 3D HOG descriptor, the angular range of ϕ was projected in an unsigned orientation range $[0°, 180°]$.

The image volume is then subdivided into a dense grid of uniformly spaced cells. However, in contrast to the one-dimensional histogram estimated for the case of HOG extraction in 2D, a 2D histogram is defined for each cell in which each histogram axis defines the angular distribution for θ or ϕ. As for the original HOG descriptor, the histogram entries for the HOG descriptor in 3D are weighted to reduce the impact of gradient orientations close to bin boundaries. Whereas the histogram entries for the 2D case only have to be split into bins of a 1D histogram, the histogram entries for the 3D case have to be split into bins of the 2D histogram, taking into consideration the binning for both orientation angles θ and ϕ.

Each histogram is then summarized in a vector by row-wise concatenation of the histogram entries describing the gradient distribution for the respective cell. After normalizing the block histogram vectors, the final HOG descriptor is obtained by concatenating these vectors.

By comparing the HOG descriptors obtained for two images, it is possible to estimate the image similarity. This can be done by a multitude of different means such as simple distance measures

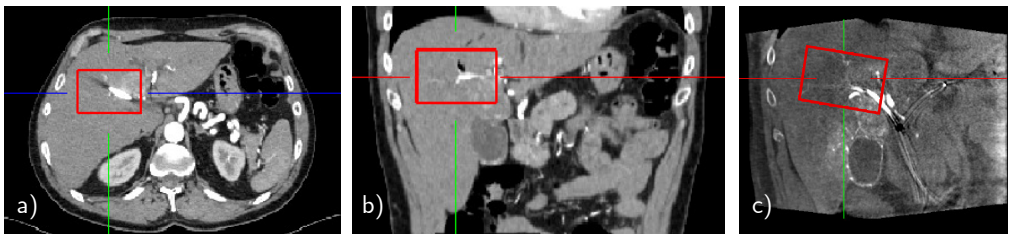


Figure 4.12: Patient data used for the evaluation of the similarity measure: a) Axial slice of the 3D-CT data; b) coronal slice of the 3D-CT data; c) coronal slice of the 3D-CBCT data after the manual preregistration. The red square in each slice indicates the ROI for the registration evaluation.

or approaches to compare probability distributions such as e.g. Kullback-Leibler-divergence [193]. Due to the simplicity of the computation, the 3D-HOG descriptor of the reference image and the 3D-HOG descriptor of the source image were compared using the Euclidean distance between both vectors for the experiments shown in this thesis. The entire workflow of this novel similarity measure is shown in Figure 4.11.

To use the HOG-based similarity measure in an image registration process, the extraction of a 3D-HOG descriptor as well as the complete HOG-based similarity metric was implemented in *C++* as module of the *ITK toolkit* [172]. To optimize computation, the reference image descriptor is only calculated once whereas the descriptor for the source image is calculated for every optimization iteration due to the geometric transform applied to the source image during every registration iteration. The variation of different parameters of the HOG extraction, such as the cell and block size, as well as the number of histogram bins, gives the possibility to optimize the performance of the HOG-based similarity measure depending on the intended application. Smaller cells or block sizes and higher number of histogram bins will increase the sensitivity of the metric to local changes in gradient distributions, but can therefore also impede the matching process of different HOG descriptors.

Experimental Setup

To validate the performance of this novel HOG-based similarity measure, an evaluation based on the methods presented in Section 4.2.4 and chapter 4.2.2 was applied. To investigate its performance in a clinical context, the patient data presented in Section 4.2.1 has been used as ground truth for the evaluation.

The focus is set to the similarity evaluation and optimization for the registration of interventional data, represented by the CBCT, to the pre-interventional CT of the patients. The results of previous experiments have shown, that this is the most challenging registration due to the high level of noise in the CBCT data sets as well as its limited field of view (see Section 5.1). Therefore, new registration approaches are required to improve the accuracy for these cases including the investigation of novel similarity measures. Taking into consideration the intended registration application during liver interventions for the treatment of HCC, the evaluation focused on the registration of a region of interest (ROI) inside the liver which was chosen individually for each of the three patients. Figure 4.12 exemplary displays the ROI (voxel size: $74.9 \times 51.6 \times 49.5\,\mathrm{mm}^3$)

selected for one of the patients.

In addition to the generation of ground truth data described in Section 4.2.1, a second preregistration step is applied consisting of an affine registration based on a mutual information image metric. The metric is implemented in a registration framework consisting of a rigid 3D transform and a Nelder-Mead downhill simplex-method as optimizer. To ensure a comparability of the results obtained for the data set of each of the three patients, the metric and optimizer parameters are not changed throughout the evaluation. The validation of the similarity metric consisted of two parts, which in turn were subdivided to study the effects of translation and rotation transformations separately.

Sampling of the Parametric Space

The first part focused on the isolated evaluation of the similarity metric by sampling of the parametric space based on the methodology presented in Section 4.2.4. Due to the nature of the CBCT and CT data, the sampling of the parametric space corresponds to the evaluation case C presented in Figure 4.9 evaluating the similarity measure in an image region with constant overlap and neglecting non-valid image region. Metric values were calculated after applying translation transformations to the source image in a range from -30 to 30 mm relative to a reference position which corresponded to the alignment determined by the preregistration. This was done for the tranlsation in x-, y- and z-direction as well as for the translation in the x-y-plane. To sample the parametric space of the metric for the case of rotational transformations, the source image was rotated in a range from $-90\,^\circ$ to $90\,^\circ$, with the rotation center set to the center of the ROI. The parametric space was sampled for the rotation around the x-y- and z-axis.

Estimation of the Capture Range

In the second part of the evaluation, the achievable registration accuracy when using the developed similarity metric was studied in terms of the capture range of the method. The registration accuracy is determined based on the evaluation methodology for rigid registration methods presented chapter 4.2.2. The method has been expanded by calculating the percentage of successful registrations per initial displacement whereas a registration is considered successful when the mTRE after registration is below 3 mm.

To enable a performance comparison of the HOG-based metric to another well-established similarity metric, the evaluation process was repeated for a normalized cross correlation (NCC) metric using the 3D-CT and 3D-CBCT data sets of one of the three patients. This patient is referred to as *patient 1* in the following. For this data, the position of the optimization minimum in parametric space found during the registration process corresponds to the transformation of the preregistration for both the HOG and the NCC metric. For the data of the other two patients, this position deviated for the NCC metric, thus preventing a fair comparison of the two similarity metrics.

4.3.2 Deep Metric Learning based on a Siamese Neural Network

Another alternative to classic multimodal similarity measures are *Deep Similarity Metrics* which rely on the training of a neural network to estimate image similarity as presented in Section 3.2. The most widely used network architectures for this task are represented by *Siamese networks*

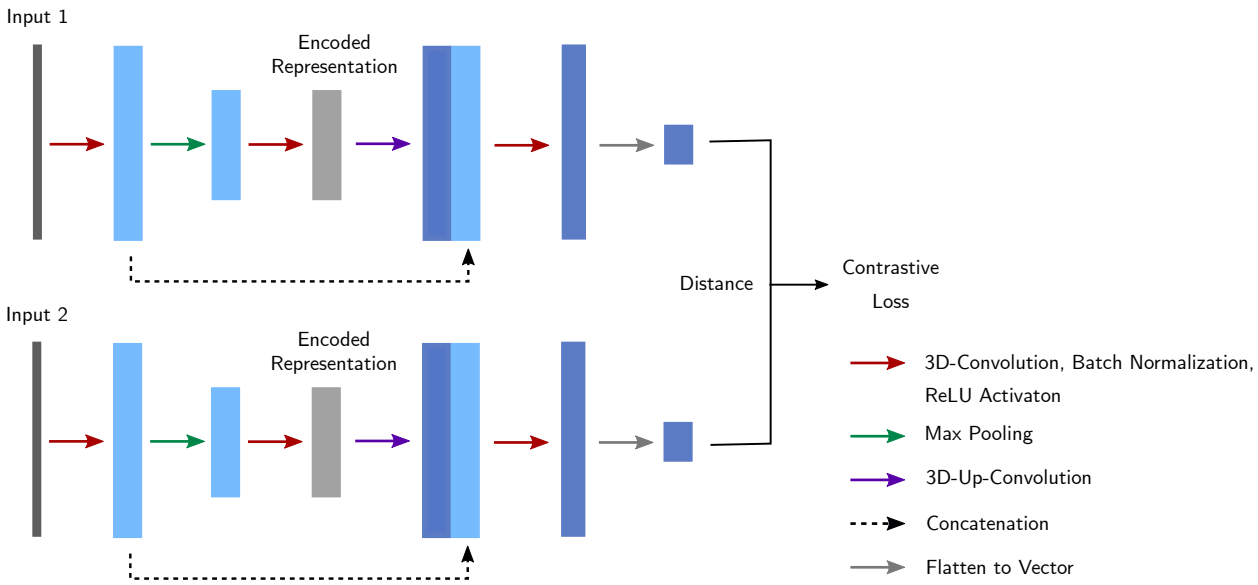


Figure 4.13: Siamese network architecture used for learning a multimodal similarity metric.

[153, 154]. The general aim is to train such a siamese network so that it is able to estimate a similarity score between patches of different medical image modalities and further analyze its behavior for different modality combinations and morphologies.

Siamese Network Architecture

For the work presented in this thesis, a siamese network is implemented using *TensorFlow* [178]. The architecture of the siamese network corresponds to a feed-forward convolutional neural network with two identical branches each of which relies on an encoder-decoder network as shown in Figure 4.13. The encoding part of the branches consists of two 3D-convolution layers. The convolutional layers use a $4 \times 4 \times 4$ sized kernel, a stride of 3 pixels and an image padding based on the duplication of the intensity values for voxels at the image edge. Each convolutional layer is followed by a batch normalization layer which adjusts and scales the activation values and a ReLU activation layer. In addition, the first convolutional layer is followed by a 3D max pooling layer with a kernel size of $2 \times 2 \times 2$ and a stride of 2 to downsample the size of the resulting feature maps. The encoding path of the network generates a feature representation of the input image in which all image characteristics that are necessary for the task of distinguishing similar and dissimilar image patches are encoded.

The decoding path consists of a 3D up-convolution layer with a kernel size of $2 \times 2 \times 2$ and a stride of 2 pixels which increases the size of the feature representation. To include additional image information in the upsampling process, the features maps which are generated during the encoding process are concatenated to the output of the up-convolutional layer. This resembles the skip connections implemented in a regular U-Net architecture. After upsampling the feature maps to the size of the original network input, the output is flattened to a vector representation of these feature maps. The element-wise distance of the output vectors obtained for both network branches then serves as input for the calculation of the contrastive loss function (see Eq. 3.11).

Table 4.5: Summary of the data characteristics used for the training of the multimodal Siamese metric models.

GT Data	Modalities	Name	Patch Sizes [px]	# Training Pairs	Margin m
IXI	T1/T2	Patches 70	70x70x70	9526	0.5
		Patches 100	100x100x100	4763	0.5
		Patches 130	130x130x130	2381	0.55
		Patches 150	150x100x150	1587	0.55
		Slices 50	238x50x188	1587	0.5
		Slices 50 Y	50x238x188	1587	0.55
XCAT	T1/CT	Patches 70	70x70x70	239	0.5
		Patches 100	100x100x100	89	0.55
		Patches 130	130x130x130	89	0.5
		Patches 150	150x100x150	29	0.5
		Slices 50	248x248x50	14	0.5
		Slices 50 Y	50x248x160	19	0.5
	CBCT/CT	Patches 70	70x70x70	239	0.6
		Patches 100	100x100x100	89	0.5
		Patches 130	130x130x130	89	0.5
		Patches 150	150x100x150	29	0.5
		Slices 50	248x248x50	14	0.5
		Slices 50 Y	50x248x160	19	0.5

Network Training

To learn the image features which are necessary for this distinction, the network has to be trained with a set of labeled, optimally registered ground truth image pairs as described in Section 3.2. As for all image processing applications that rely on registered multimodal ground truth data, this requirement represents a major challenge since this type of multimodal data is only scarcely available. Therefore, the training data used for the training of the siamese network is extracted from different ground truth data sets presented in Section 4.2.1. Only the ground truth consisting of real patient data (Sec. 4.2.1) is excluded, since the preregistration of these data pairs is not accurate enough to extract training patches for the siamese network.

The siamese network is then trained with image patches of different modality combinations and also different types of morphology to study the applicability of the learned Siamese Deep Metric (SDM) to various data sets. Moreover, different spatial geometries of the training patches are investigated to derive training characteristics of the metric. The first data set used for training the network corresponds to the registered T1 and T2 scans of the brain derived from the IXI data set (Sec. 4.2.1). The intensity values of the registered image volumes are normalized to a range of $[-1, 1]$ to ensure a stable learning process by limiting the distribution range of feature values in the images. Next, 3D subvolumes are extracted from the native scans at the same image positions in both modalities and then used as training patches. To further understand the impact of the geometry of the training patches on the performance of the learned Siamese metric, the patch size is varied according to the sizes listed in Table 4.5. This table also includes the names employed for distinguishing the input sizes in the following chapters. To ensure that no patches are used for the training of the network that only include background information, the sampling of subvolumes is

limited to the region of the brain.

The benefits of the extraction of subvolumes are twofold: the extraction yields a major benefit by increasing the amount of available image pairs for training of the network. For the IXI data set, the subsampling process yields multiple hundred registered image pairs whereas the correct number for each patch size is listed in Table 4.5. In addition, the use of smaller image patches reduces the computational memory required for the training of the siamese network. Nevertheless, due to the fact that the siamese network is implemented for 3D image processing, the number of patches that can be processed simultaneously is still limited, leading to the employment of a batch size of 6 during training of the network.

The sampling procedure results in a set of registered intermodal T1 and T2 MRI image patches that are labeled as 'similar'. To generate dissimilar image pairs, T1 MRI patches are paired with random T2 MRI patches and labeled as 'dissimilar'. To avoid overfitting, up to 3% random noise is added to the T2 patches of each similar and dissimilar input pair. For every training process, 75% of these image pairs are used as training data and the remaining 25% serve as validation data.

The second data set relies on the synthetic multimodal image data of the abdomen that is generated based on the digital XCAT phantom using image-to-image translation as described in Section 4.2.1. The siamese network is trained for the data combinations CBCT/CT and T1/CT, since these represent the modality combinations which are most relevant for the clinical context of this thesis. The preprocessing of this data is very similar to the preprocessing of the IXI data sets, including an intensity normalization to the range of [-1,1] and the extraction of 3D subvolumes in the abdomen at fixed locations in each image modality with various patch sizes listed in Table 4.5. As for the IXI data set, these extracted patches serve as 'similar' image pairs and a shuffling of these pairs is applied to generate 'dissimilar' image pairs. Then resulting image pairs are again divided, so 75% serve as training data and 25% as validation data.

So in summary, siamese network models are trained for three separate application cases, including the distinction between similar and dissimilar subvolumes of

- T1- and T2 MRI data of the brain,
- synthetic CBCT and CT scans of the abdomen,
- as well as synthetic T1 and CT scans of the abdomen.

For each application, the training is performed by optimizing the contrastive loss function given in Eq. 3.11 using the labeled image patches and an Adam Optimizer with a learning rate of $lr = 2{\cdot}10^{-4}$ over 30 epochs and batch size of 6. It has to be noted, that the optimal margin value m in the contrastive loss function varies according to the image data and image size. In general, this margin value designates a distance value and training pairs with a distance larger than this value will not be considered for the calculation of the loss. Therefore, this value contributes as a regularization value forcing the network to learn small deviations to distinguish between images that are labeled 'dissimilar' but still display similar structural features. If it is chosen too large, highly dissimilar patches still contribute to the loss, if it is chosen too small, the siamese network losses general

validity. The optimal value for m for each set of training data is empirically determined using only values that yield a smooth parametric cone and listed in Table 4.5. Since the Euclidean distance used for the calculation of the contrastive loss is normalized using the size of the employed training patches, the values for m are very similar for the training with different patch sizes ranging between 0.5 to 0.6.

Experiments

Similar to the first part of the evaluation of the HOG metric, the performance of the different Siamese metric models is evaluated based on the sampling of the parametric space for translation and rotation transformations. The sampling is performed according to the three metric evaluation cases presented in Section 4.2.4. For evaluation type B, the offset of the position of optimal image alignment is chosen as 10 mm in the investigated translation direction for translation transformations and as rotation of 10° around the studied rotation axis for rotation transformations.

To sample the parametric space, the translational image transformations are chosen between -30 mm and 30 mm and the rotational transformations between $-30°$ and 30°. The sampling of the parametric space for translation transformations includes the translation of one of the images along the x-,y- and z-direction of the coordinate system. These axis correspond to the sagittal, coronal and axial-direction of the image subject (depending on the data, either the brain or the abdomen). To include a translation direction that does not correspond to one of the main axes, the parametric space is also sampled for translations along the axis (1,1,0). As for rotation transformations, the moving image is rotated around the x-, y- and z-direction as well as the diagonal axis (1,1,1). The performance of the different metric models is evaluated on the same image modality combinations that are used for training. Since the siamese network that is used for learning image similarity is exclusively composed of convolutional layers, the network is able to process image data of various size and not just the input sizes it has been trained for. To demonstrate this ability, the image data used for all metric evaluation experiments, corresponds to axial 3D slices that are extracted from the original data sets that is used to generate the multimodal training patches. Concerning the IXI data, these slices have a size of $248 \times 160 \times 48$ pixel whereas the XCAT slices have a size of $248 \times 248 \times 80$ pixel.

The evaluation methodology is implemented in *python* relying on *Keras* for loading and applying the pretrained model of the SDM. The geometric transformation for the sampling of the parametric space is realized using a *Spatial Transformer Network* (STN) [171]. This STN outputs the geometrically warped moving image as well as a binary image mask determined during the transformation covering the image regions in which valid image information is stored after the resampling process. During the estimation of metric values, this mask is used so that only image values in the geometrically warped image that are located in the non-zero regions of the image mask are considered during the calculation of the metric value. This ensures that only valid image regions contribute to the metric evaluation.

As for the HOG metric, the results of the parametric sampling are additionally compared to the results obtained using traditional similarity measures. Since the Siamese Deep Metric is mainly

investigated for multimodal image similarity estimation, the comparison is performed using a multimodal similarity measures only and not the NCC metric which was used to benchmark the HOG-metric.

As first alternative multimodal metric, an NGF loss relying on the alignment of image gradients (see Eq. 3.8) is implemented in *Keras*. And advantage of the NGF loss in comparison to the SDM is the fact that no model has to be trained and loaded for the similarity estimation. However, the Siamese metric is expected to yield a higher sensitivity for multimodal image data, since it is directly trained on the intended data sets.

Moreover, the results of the parametric sampling are additionally compared to the results obtained using the Advanced Mattes Mutual Information (AMMI) similarity metric implemented in the *ITK* image processing toolkit. The metric is used twice, with and without the application of an image mask covering only valid image regions. For reasons of simplicity, the geometric transformation applied for this part of the evaluation was performed using *ITK* instead of the STN, although both methods result in the same image warping.

Moreover, the performance of the SDM trained with the synthetic abdominal data is evaluated on a test set of synthetic CBCT/CT and T1/CT image pairs as well as on real patient data represented by the preregistered abdominal CBCT and CT scans of patients presented in Section 4.2.1. The idea is to train the metric model using synthetic image data and then apply the learned SDM on patient data to investigate its transferability. A successful transfer could highly increase the usability of the novel similarity metric in a clinical setting. The transfer is facilitated by the fact, that both data types (synthetic and patient data) cover the same abdominal region and the patient CBCT/CT data displays a very similar grey value distribution as the synthetic CBCT/CT data. Since the CycleGAN used for the generation of the synthetic ground truth data (see Section 4.2.1) is trained with T1 MRI data based on a different image acquisition sequence as the T1 MRI of patient data, both T1 MRI display different grey value distributions. Thus, a transfer of the SDM trained with synthetic T1 MRI to real patient data is not appropriate, and the evaluation of the SDM on real patient data is restricted to CBCT/CT.

4.4 End-to-End Image Registration Learning

In previous sections, improvements for single components along the pipeline of an image registration process were presented. Since image registration is an iterative process, the overall estimation of optimal transformation parameters can still be time consuming. The aim of end-to-end registration learning is to estimate the transformation parameters directly in a single step. Hence, aside from *Metric Learning*, the power of Deep Learning is also leveraged in the field of medical image registration by training neural networks to learn an entire registration process. A widely used example for a fast, self-learning convolutional neural network is the open-source VoxelMorph network which was presented in more detail in Section 3.3. This network is capable of performing monomodal deformable registration tasks while achieving state-of-the-art registration accuracy. Yet, the network enables a reduction of the required computation time several orders of magnitude compared to registration methods relying on traditional image processing.

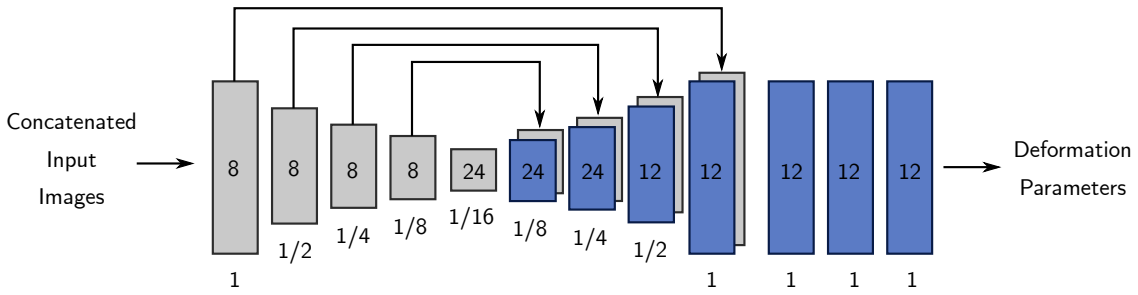


Figure 4.14: Convolutional U-Net architecture building the basis for the VoxelMorph network used for the experiments presented in this thesis. The input to this network is represented by a 2-channel 3D volume consisting of the concatenated reference and source image. The network output serves as transform parameter input to the Spatial Transformer Network. Each rectangle represents a 3D volume and the arrows the skip connections used to concatenate encoder and decoder feature maps.

Within the frame of this thesis, the unsupervised version of the VoxelMorph network presented in [6] and referred to as 'original' VoxelMorph network in the following, has been modified. The modifications include alterations in terms of geometric transformations by allowing the computation of affine geometric transformations in addition to deformable image transformations. Moreover, novel similarity metrics and loss functions have been integrated and investigated to extent the VoxelMorph network for the application on multimodal image registration tasks.

4.4.1 Extension of the Classical VoxelMorph Network Architecture

The original code of the VoxelMorph network is implemented using *Keras* with a *TensorFlow* backend and made publicly available by the authors [194]. The extension of the network is based on the original code and all alterations of the network are also implemented using *Keras* with a *TensorFlow* backend.

For the implementation of a multimodal VoxelMorph network, the general encoding-decoding architecture is kept similar to the VoxelMorph-1 network presented in [170]. However, the filter number of the convolutional layers is adjusted for the task at hand, so that the network consists of three convolutional layers with 8 filters and one layer with 24 filters for the encoding part of the network. The decoding part is constructed of two upsampling convolutional layers with 24 filters and four with 12 filters as shown in Figure 4.14.

Geometric Transforms

The original VoxelMorph network is designed to learn a deformable image registration. This type of registration method relies on the computation of a deformable geometric transformation that is differentiable and invertible, which leads to a smooth deformation field preserving image-topology [168, 169]. In general, deformable image registration methods are very flexible and useful to model organ deformations in medical image processing (see Section 2.3.2). However, these methods are designed to capture local image motion rather than larger global image displacements. Thus, deformable image registration methods are mostly applied on pre-aligned image data sets to correct for small structure deformations.

To make the VoxelMorph registration applicable for the registration of larger global image displacements, the first extension of the original network included the implementation of affine registration learning. The original network is designed so that the encoder-decoder block computes a 3D volume whose size is given as the image size with 3 channels. Each channel can be interpreted as the displacement vector entry for one of the three spatial dimensions x, y and z. This output then serves as input to the Spatial Transformer Network [171], which uses the volume elements as parameters for estimating a deformation field. This resulting 3D deformation field is then applied to the source image to compute the warped image using linear interpolation. In addition to the warped source image, the STN also outputs the applied deformation field.

To enable the network to learn affine registration tasks, a 3D global average pooling layer is added to the encoder-decoder block. This additional layer estimates the average for each of the 12 feature maps (see Figure 4.14), resulting in 12 single values. These values correspond to the 12 parameter required to define a 3D affine transformation (9 parameters defining the rotation matrix and scaling + 3 translation parameters). So instead of a 3D deformation field, the input to the STN is represented by 12 parameters which the STN arranges to a two-dimensional 3×4 affine transformation matrix. The STN calculates a deformation field based on this matrix which is applied to create the warped source image.

This extension of the architecture enables to choose between training the VoxelMorph network for affine or deformable registration tasks.

Application of Image Masks

In addition to the extension of the network for learning affine image transformations, the STN is modified so that it not only yields the warped source image and the 3D deformation field used to create this warping, but also an image mask. This binary mask is computed during the transformation and has zero values in image regions containing non-valid image information that are created during the image resampling process (see Section 4.2.4). It is concatenated to the warped source image and the deformation field, so that the STN outputs a 5-channel tensor. This mask allows to take only valid image regions into account for the metric evaluation during the registration learning process. Ideally, the application of an image mask is suggested to increase the registration accuracy of the network, since non-valid image regions are neglected for the learning process.

Alternative Loss Functions

Another extension of the network is represented by integrating alternative loss functions for the training of the network. The loss function of the VoxelMorph network is a combination of a similarity loss L_{sim} that compares the image appearance of the reference and the warped source image and a loss designated to enforce a smooth deformation field L_{smooth}. Both losses are multiplied by a weighting factor (by default set to 1 in the original VoxelMorph network) and added up to a complete loss function. By changing the weight parameter, one of the two losses can either be set to a higher importance or completely neglected by setting the value to 0.

In the original unsupervised version of the VoxelMorph network, L_{sim} can be set to either a negative cross-correlation loss or to a mean-squared error loss. Both of these similarity measures perform best for image pairs that display a similar intensity distribution and are therefore mostly used for monomodal registration purposes. Moreover, both metrics are implemented without the consideration of an image mask. This means that non-valid image regions are taken into account for the metric evaluation during registration learning in case of larger global image displacements. This possibly decreases registration accuracy. Within the frame of this thesis, new loss functions L_{sim} are implemented, such as:

- Mean-Squared Error Loss with Image Mask:
  A first addition to the original loss functions is realized by implementing a mean-squared error (MSE) loss function that relies on the use of an image mask. Before returning the MSE metric value, the distance tensor is multiplied with the binary mask that is calculated during the geometric transformation by the STN, so that non-valid image regions are neglected during similarity estimation. Moreover, the mask is used to normalize the loss value. Therefore, the mask entries are added up to determine the number of valid image voxel. The loss value is then divided by this number to increase robustness of the loss with respect to the size of the overlap between fixed image and warped source image.

To extend the possible applications of the VoxelMorph network to multimodal registration tasks, additional loss functions are implemented bearing the potential to be applied on and multimodal image data:

- Siamese Metric Loss (unsupervised):
  The first multimodal loss function is given by the integration of the pretrained models of the siamese network, described in further detail in the previous chapter 4.3.2, in the VoxelMorph network. These models are trained to evaluate image similarity between different modalities. By incorporating the graph of the siamese network in the VoxelMorph network and loading the pretrained siamese network weights, the SDM can be used to evaluate image similarity during the training of the VoxelMorph network. It has to be noted that the weights of the siamese network branch have to be excluded from the pool of trainable network parameters to enable the use of the pretrained models. By using the SDM in the VoxelMorph network, the network is able to learn the registration of multimodal datasets. However, since the SDM was trained for specific modality combinations and data sets, the siamese loss is only applicable for the registration of these data sets. During the metric evaluation, non-valid image regions are masked out using the binary image mask.

- Deformation Field Loss (supervised):
  An alternative supervised learning method is implemented which is based on the comparison between two deformation fields. Using optimally registered image pairs as input data to the network, one of these images is geometrically altered by application of a transformation with known transform parameters using *SimpleITK*. These transformation parameters can be used to estimate a 3D deformation field. By penalizing differences between this known field

to the deformation field provided by the network, a novel loss function is created to train the network. This deformation field loss (DFL) function is suspected to be applicable for mono-and multimodal registration learning. However, this approach relies on ground truth data in form of optimally registered images or a 3D ground truth deformation field, and is therefore no longer considered as unsupervised.

4.4.2 Training VoxelMorph Models

Different VoxelMorph registration models are trained using the original as well as the modified network architectures. To cover monomodal and multimodal registration learning, different models are trained separately for both applications and various input data pairs.

Training Data

These input data pairs are extracted from the ground truth data described in more detail in Section 4.2.1. The first data set which is used to train different VoxelMorph models corresponds to the registered T1 and T2 brain MRI scans of the IXI data set. The second image data set is represented by the synthetic abdominal CT, CBCT and MRI scans that were simulated for five positions of the respiratory cycle and that are registered for each position. Due to the respiratory motion, the five data triplets are characterized by slight variations in morphology. To simplify the registration learning process, all images are resampled to a uniform physical spacing corresponding to $1 \times 1 \times 1\,\mathrm{mm}^3$ for both data sets.

Since both data sets represent ground truth data that consists of optimally registered image pairs, the image chosen as moving image is geometrically transformed for all application cases to create an actual registration task. The aim is to create a realistic registration task that corresponds to the scenario denoted as *Case B* in Section 4.2.4. This transformation is applied directly after loading an image pair for training the network. The transformation is realized by application of a rigid transform using *SimpleITK*. The translation direction is randomly chosen and the range is set between -30 and $30\,\mathrm{mm}$. The rotation angle is also randomly chosen between -15 and $15°$ with a random rotation axis tilted by $10°$ maximum relative to the axial direction of the skull or abdomen. The geometric transformation not only generates appropriate image data for registration learning, but also yields a ground truth deformation field which is deviated from the applied transform parameters using *SimpleITK*. This ground truth deformation field can be used for the calculation of the deformation field loss described in the previous chapter.

Due to computational limitations in terms of memory usage, the network is trained with 3D subvolumes corresponding to 3D axial slices of the image volume instead of using the whole image. These slices are extracted at variable axial positions between 80 – $100\,\mathrm{px}$ from the original 3D volumes with spatial dimensions of $160 \times 256 \times 80$ for the IXI data and $320 \times 320 \times 64$ for the synthetic multimodal data.

For monomodal registration learning, the first training data consists of T1 MRI brain scans of the IXI data as fixed and moving image. In addition to the geometrical transformation of the moving image, random noise up to 6% is added to the image to increase robustness and avoid overfitting of the model. To train and test the monomodal registration learning on multiple data sets, the

Table 4.6: List of monomodal and multimodal registration learning models and their respective characteristics.

# Model	Registration Type	GT Data	Modalities	Transform	Loss Function
1	Monomodal	IXI	T1/T1	affine	MSE w/o mask
2					MSE
3					DFL
4				deformable	MSE w/o mask
5					MSE
6					DFL
7		XCAT	CT/CT	affine	MSE w/o mask
8					MSE
9					DFL
10				deformable	MSE w/o mask
11					MSE
12					DFL
13	Multimodal	IXI	T1/T2	affine	DFL
14					SDM
15				deformable	DFL
16					SDM
17		XCAT	CBCT/CT	affine	DFL
18					SDM
19				deformable	DFL
20					SDM
21			T1/CT	affine	DFL
22					SDM
23				deformable	DFL
24					SDM

synthetic abdominal CT scans serve as second training data set. The same preprocessing is applied as for the IXI data including the addition of noise and the application of a geometric transformation to the moving image. Both data sets are used to train multiple monomodal VoxelMorph registration learning models, varying different parameters such as the loss function, the application of an image mask during similarity estimation or the type of geometric transformation. A list of all trained monomodal registration models is given in Table 4.6, listed as model number 1 to 12.

For multimodal registration learning based on the VoxelMorph network, the T1 and T2 MRI scans of the IXI data set as well as the synthetic abdominal CBCT, CT and MRI scans are used to generate training image pairs. For the network training based on the IXI data set, the T1 MRI serves as fixed image and the T2 MRI as moving image. Since both images already display variations in the intensity distribution, no additional noise is added to the images. Moreover, the synthetic abdominal data is used as training data whereas the focus is set on the multimodal registration learning for two different modality combinations: CBCT to CT and T1 MRI to CT. The CT scan serves as fixed image for all experiments. As for the multimodal IXI data pairs, no noise is added to the images. With these multimodal data pairs, different registration learning models are trained

that are listed in Table 4.6 from model number 13 to model number 24.

Training Parameters
The models are trained using a batch-by-batch data generation based on *Keras*-intern functions. This guarantees an optimal use of memory exploitation due to real-time feeding of the model with input data. The different networks are trained using a total of 500 training epochs for the IXI data. Due to the limited number of XCAT data pairs corresponding to five per modality combination, an increased number of 5000 epochs was used for training models using the synthetic ground truth data. By increasing the number of epochs, the number of training pairs is artificially augmented, since the z-position of the extraction of the 3D slices which are used for training is varied between 80 - 100 px for every epoch and thus the training pairs are different for every epoch. Moreover, the parameters of the ground truth deformation field are also varied for every training run which additionally leads to an effective augmentation of training data.

As an additional alteration to the original VoxelMorph code, an adaptive learning rate is used for all trainings implemented by using the *Keras*-intern function *Callbacks.ReduceLROnPlateau*. This function reduces the initial learning rate by a defined factor once the learning process stagnates and no decrement of the loss is captured for a specified number of epochs. For both data sets, an initial learning rate of $lr = 2e^{-4}$ is chosen, which is reduced by a factor of 0.8 after 15 epochs without a decrement of the loss function up to a minimum learning rate of $lr = 1e^{-8}$for the IXI data set and after 300 epochs for the XCAT data. Due to memory limitations, a batch size of 1 was employed for all model trainings.

Since the total loss function corresponds to the sum $L_{VM} = a \cdot L_{sim} + b \cdot L_{smooth}$, the weighting factors a, b for each part of the total loss function are adjusted according to the investigated loss L_{sim}. This is necessary due to the fact that the magnitude of the loss values highly varies for different loss functions L_{sim}. The authors of the original VoxelMorph network recommend a ratio of $\frac{a}{b} \approx 10$. Therefore, all models are trained with a similar ratio. The weighting factors are adjusted empirically for every model to ensure that the contribution of both losses to the total loss is the same for all training runs.

4.4.3 Experiments

To characterize and evaluate the performance of the original VoxelMorph network as well as the adjusted and extended versions for mono-and multimodal registrations, the trained network models are linked to the evaluation methodology presented in Section 4.2.2. The aim is to establish a better understanding of the different variants of the VoxelMorph network by determining their respective capture range in terms of registration accuracy. All model numbers listed in the following refer to the numbers listed in Table 4.6.

Since the MSE loss relies on a similar intensity distribution of the input images, all experiments relying on registration models using an MSE loss are performed for monomodal data only.

- **MSE loss function with and without a mask**
  Ideally, the use of a binary image mask during metric evaluation while registration learning

is suggested to increase the accuracy of the registration, since non-valid image regions are neglected. Therefore, the performance of the affine and deformable network versions trained using a MSE loss without a binary mask is compared to the performance of these network trained using a MSE loss with a binary mask.
Models investigated: 1 & 2, 4 & 5, 7 & 8 and 10 & 11

The following experiments are valid for monomodal and multimodal registration models. However, not all studies are carried out for both registration types (or all available data sets) in cases where one application is enough to demonstrate a general feasibility.

- **Affine vs. deformable registration learning**
  In general, deformable image registration techniques are applied to correct for small local structure deformations and affine registration methods for larger global displacements. Thus, affine registration methods often provide a larger capture range as deformable methods. The performance of the original deformable VoxelMorph network is therefore compared to the performance of the extended affine VoxelMorph network. This comparison is performed for all models based on mono- and multimodal loss functions.
  Models investigated: 1-3 vs. 4-6, 7-9 vs. 10-12, 13-14 vs. 15-16, 17-18 vs. 19-20, 21-22 vs. 23-24

- **Deformation field loss function**
  Supervised learning optimizes the model performance based on ground truth data. In general, this leads to the best approximation of the relationship between input and desired output data. This is further studied, by comparing the performance of the registration methods trained using the supervised deformation field loss to unsupervised methods. Since the deformation field loss is applicable for mono- and multimodal data, both registration types are investigated.
  Models investigated: 3, 6, 9, 12, 13, 15, 17, 19, 21, 23

Since the Siamese metric is trained for specific modality combinations and data sets, the experiments to study the performance of registration models trained with the Siamese metric loss are only applied on multimodal data. However, the metric could potentially also be trained and applied for monomodal registration tasks.

- **Siamese metric loss for multimodal registrations**
  As an alternative multimodal metric, the SDM has been trained for the different modality combinations and incorporated in the VoxelMorph network. The registration performance has been investigated for affine as well as deformable multimodal registration models.
  Models investigated: 14, 16, 18, 20, 22 and 24

These studies aim to investigate the behavior and applicability of the classical VoxelMorph registration network as well as the extended variants for monomodal and multimodal registration tasks. Since the alterations to the original VoxelMorph network proposed in this chapter aim to extend its application to registration tasks that require the correction of global image misalignments, the

evaluation is restricted to rigid registration tasks and deformable registration is not considered in this context. Hence, the evaluation is performed using the evaluation methodology for linear registration methods presented in Section 4.2.2. Although the models are only trained to correct translations up to 30 mm, the evaluation is performed for translations up to 60 mm to characterize the capture range of the method.

Since the VoxelMorph models are trained on MRI brain scans and synthetic abdominal data, this is also the data used in the evaluation methodology. Therefore, the evaluation presented in Section 4.2.2 is extended for these data sets. The preprocessing of the brain data and the synthetic data is similar to the preprocessing of the patient data described in Section 4.2.2 and only shortly summarized here. Since the data corresponds to already registered ground truth data, the preprocessing does not include a preregistration but only the segmentation via thresholding of target structures. Concerning the IXI data set, the target structure is represented by the soft tissue of the brain. The resulting segmentation is used to distribute equally spaced points inside this target region which are used for the calculation of the TRE. For the synthetic abdominal data, a liver segmentation is extracted as binary mask. The intrinsic landmarks in the liver which are directly provided by the digital phantom that is used to generate the synthetic data sets as presented in Section 4.2.1 serve as target points for the calculation of the TRE. Moreover, the segmentation mask of both data sets are used to limit the image region in which the registration accuracy is estimated to the structures of interest.

Similar to the evaluation of the Siamese Deep Metric presented in Section 4.3.2, the evaluation of the novel VoxelMorph registration models also includes the transfer of the models trained with synthetic abdominal image data to real patient data. The aim is to further characterize the performance of the models and investigate their transferability to novel different data sets. For these experiments, exempalry chosen data pairs of the abdominal CT and CBCT scans of the preregistered ground truth data presented in Section 4.2.1 serve as patient data. The evaluation includes the monomodal registration of CT to CT data as well as the multimodal registration of CBCT to CT data using the models trained for the respective modality combinations.

5

Results

In this chapter, the results of the different research projects introduced in Section 4 are presented. The chapter is divided into four parts. The first and second part correspond to the results obtained for the evaluation of linear and nonlinear multimodal registration methods, respectively. In the third part, the performance of the two novel image similarity measures based on *Histograms of Oriented Gradients* as well as a *Siamese neural network* is presented. The fourth and last part includes the results obtained for the end-to-end registration based on a neural network.

5.1 Evaluation of Linear Image Registration

In this chapter, the results obtained for the evaluation of linear registration methods are presented. Since the aim of this thesis is the optimization and evaluation of registration methods for abdominal scans in the context of interventional procedures such as e.g. biopsies, the evaluation focuses on an optimized registration of the liver tissue.

It has to be noted that the experiments were performed for all three patients, but for reasons of clarity of presentation only the results for one of the patients are shown in most of the figures. Unless stated otherwise, the results presented for this patient are similar for the data of all three patients. The results were published in the proceedings of the international conference *SPIE Medical Imaging 2019* [180].

5.1.1 Initialization

The first part of the evaluation focused on the influence of initialization methods on the image registration process. Three different initialization methods are compared:

1. The initialization by superposition of the geometrical image centers.
2. The initialization by superposition of the geometrical image center of the reference image (CT) and the geometric liver center of the source images.
3. The initialization by superposition of the liver centers.

Table 5.1: Median TRE (in mm) of the manually chosen landmarks after performing different approaches for a registration initialization for data sets of three patients.

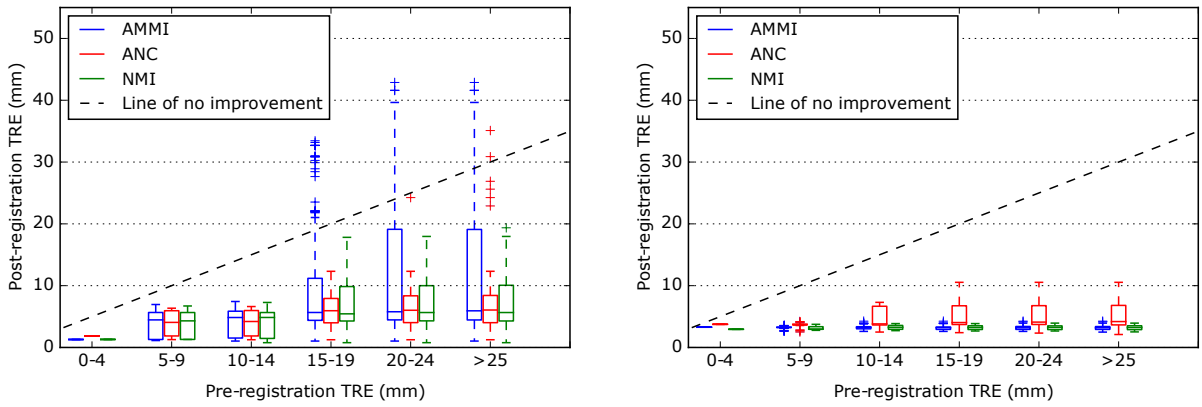
Modalities	Initialization	Pat 1	Pat 2	Pat 3
CBCT to CT	Geometrical Center	83.49	84.48	85.85
	Liver Center/Geometrical Center	26.52	37.88	25.50
	Liver Center	20.46	25.21	13.07
T1 to CT	Geometrical Center	9.75	33.02	35.16
	Liver Center/Geometrical Center	111.56	74.68	96.35
	Liver Center	19.86	56.84	27.79
T2 to CT	Geometrical Center	11.63	43.11	19.40
	Liver Center/Geometrical Center	107.51	71.21	98.23
	Liver Center	17.50	53.60	29.52

The different initialization methods are compared in terms of the median Target Registration Error (mTRE) using the manually chosen landmarks. Since the initialization methods are compared using non-preprocessed image data, no ground truth artificial landmarks are available for this part of the evaluation. However, the results presented in the following all rely on artificial landmarks, since only a limited number of manually chosen landmarks is available.

The results of the different initialization approaches are listed in Table 5.1. In general, these TRE values correspond to the physical distance that would have to be corrected by a registration algorithm after an initialization has been performed. For the registration of CBCT to CT, it can be stated that the initialization based on the superposition of the liver centers in the CBCT and the CT images yields the lowest mTRE of the manual landmarks. Thus, this method outperforms the other two approaches based on alignment accuracy. This is reasonable, since one of the main obstacles when registering CBCT to CT data is the fact that the CBCT covers a smaller field of view than the CT which is often restricted to the region of interested during the intervention. Concerning the data used for these experiments, the CBCT is acquired during a liver intervention, so that the liver is located in the geometric center of the CBCT. Since the CT covers the complete abdominal region, an initialization based on the superposition of geometric centers as well as liver center/geometric center still results in an offset of the images, as can be seen in the data listed in Table 5.1.

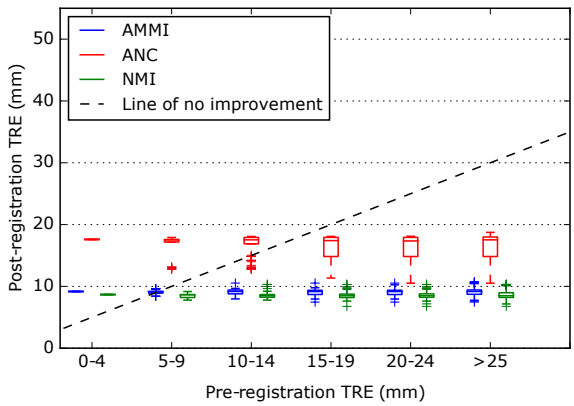
For the registration of T1 and T2 to CT, the best results in terms of the lowest mTRE are achieved based on an initialization by superposing the geometric image centers. However, the superposition of liver centers yields only a minimally increased TRE. This is due to the fact that both, the CT as well as the MRI cover the complete abdominal region and therefore the same organ structures are located in similar image regions.

The mTRE after initialization gives a first impression of the required capture range of the registration methods to obtain an accurate image alignment. The smaller the TRE, the smaller is the displacement which has to be corrected by the registration algorithm. By applying an initialization before registration, the registration accuracy can be increased.

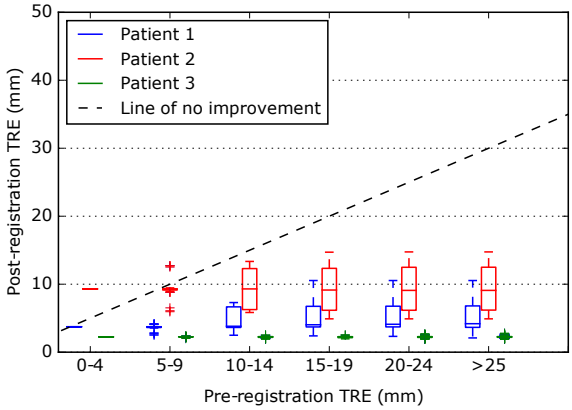


(a) CBCT to CT (with liver mask)

(b) T1 to CT (without mask)



(c) T2 to CT (without mask)



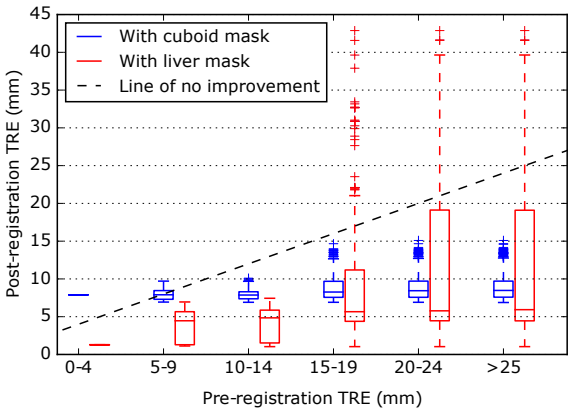
(d) T1 to CT (ANC, without mask)

Figure 5.1: **Similarity Metrics**: Results obtained for the registration of (a) CBCT to CT, (b) T1 to CT and (c) T2 to CT exemplary shown for one of the three patients using different similarity metrics, as well as (d) the comparison of the results obtained for the registration of T1 to CT using the ANC similarity metric for all three patients.

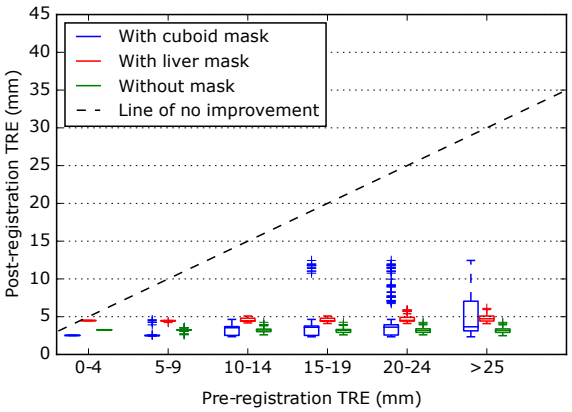
5.1.2 Similarity Metrics

The registration accuracy in the following relies on the calculation of the mTRE using the artificial landmarks. The results obtained by applying different similarity metrics for the registration of CBCT to CT, T1 to CT and T2 to CT are shown in Figure 5.1. All similarity metrics result in a very similar post-registration TRE for small initial TREs in a range from $0 - 4\,\text{mm}$. This implies that all methods register to positions close to one another, hence indicating a sufficient accuracy of the landmark-based ground truth registration.

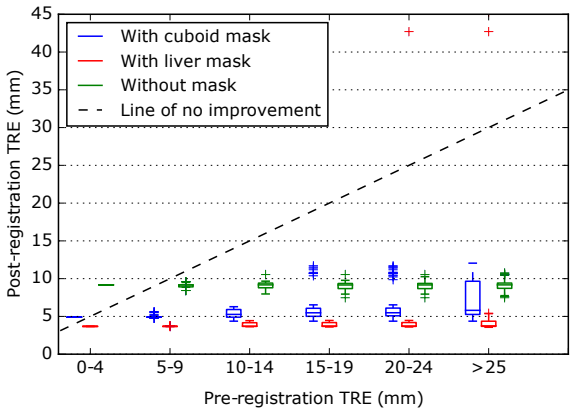
For the registration of CBCT to CT data of all patients, the application of all three similarity metrics resulted in similar post-registration accuracies with ANC and NMI slightly outperforming the AMMI metric for all three patients by yielding a lower variance (Figure 5.1a). Up to initial displacements of the liver around 14 mm, all registration methods resulted in an improvement of the image alignment. For larger initial deviations, the variance increases as a result of registrations which yield a higher TRE than before registration. Taking into consideration the results for a transformation initialization shown in Table 5.1, an improved initial image alignment superposing



(a) CBCT to CT (AMMI)



(b) T1 to CT (AMMI)



(c) T2 to CT (AMMI)

Figure 5.2: **Application of Masks**: Results exemplary shown for one of the patients obtained for the registration of (a) CBCT to CT using the liver segmentation and a cuboid as fixed image mask; (b) T1 MRI to CT and (c) T2 MRI to CT without mask, using the liver segmentation and using the cuboid segmentation as fixed image mask during registration.

the center of the liver could be beneficial. Moreover, it has to be noted that for all patients the liver segmentation served as fixed image mask during registration of CBCT to CT to prevent the registration from failure. This behavior is further analyzed in the following section 5.1.3.

Concerning the multimodal registration of T1 to CT (Figure 5.1b), the ANC metric, which in general is considered a suitable metric for monomodal applications only, since it relies on a direct relation between the grey values in the images to register, achieved similar results to the MI-based methods even resulting in a lower variance for large pre-registration TRE. To demonstrate that this finding is not patient-specific, the results for all three patients are shown in Figure 5.1d. For the multimodal registration of T2 MRI to CT (Figure 5.1c), ANC could not compete with MI-based methods, even increasing the TRE after registration. For all three patients and all modality combinations, NMI yields a low post-registration TRE outperforming the other two similarity measures.

5.1.3 Application of Masks

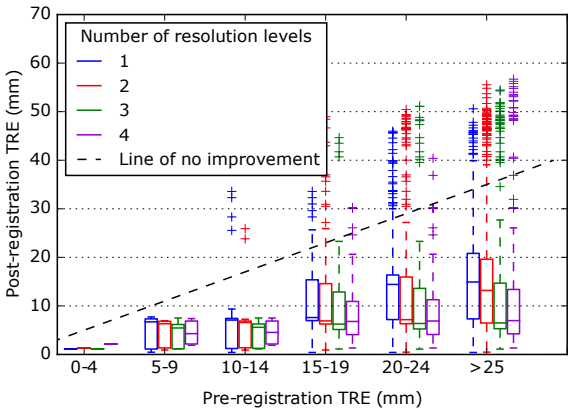
For the registration of the CBCT data sets (moving images) to CT data (fixed images) of all three patients, the introduction of a fixed image mask was required to prevent the registration from failure and achieve a sufficient registration accuracy. To study the impact of mask characteristics on the registration accuracy, two masks, the liver segmentation of the CT data obtained during image preprocessing and a simple cuboid covering the liver region, were used as binary masks for the fixed image. The results obtained for the registration of CBCT to CT are shown in Figure 5.2a. Relying on the default setting of the parameter map, AMMI was used as similarity metric. The boxplots show that the median registration accuracy increases for the more complex liver mask. However, for image displacements larger than 15 mm, the increased variance for the registration using the liver mask implies a decreasing robustness of the method.

Additionally, a comparison of the results obtained for the multimodal registration of T1 and T2 to CT for the same patient without using a mask, with the liver mask and with the cuboid mask as fixed image mask are shown in Figure 5.2b and Figure 5.2c, respectively. Concerning the registration of T1 to CT, the application of masks resulted in a similar TRE than the registration without mask. However, the highest median post-registration TRE was obtained using the liver mask. In this case, the mask reduces the image information to the homogenous region of the liver, neglecting the information of the surrounding tissue which increases the difficulty of the registration. For the registration of T2 to CT, the introduction of a mask seemed to increase the attainable registration accuracy and the median TREs for both masks were very similar.

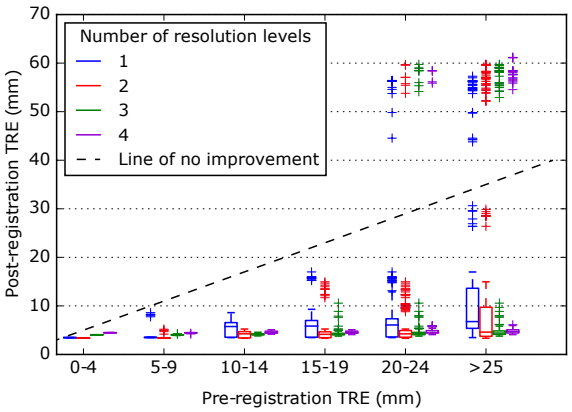
The results show that a fixed image mask improved the registration outcome for the registration of CBCT to CT and T2 to CT, whereas the application of a cuboid fixed image mask seemed to yield stable results for all modality combinations for initial displacements smaller than 25 mm.

5.1.4 Number of Resolution Levels

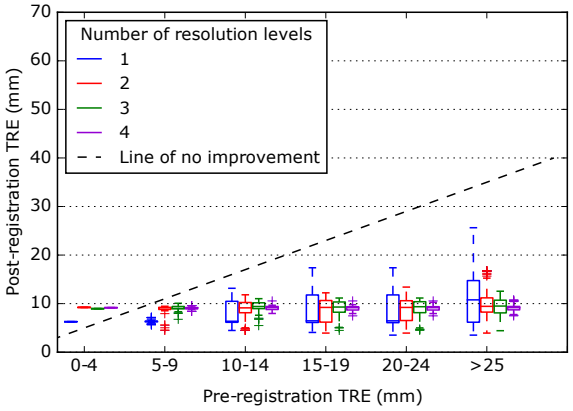
The experiments revealed that the multimodal registration of the interventional CBCT to preinterventional CT data is an especially challenging task. This is caused by the high noise level of the CBCT data and its limited field of view that only covers the anatomical region close to the liver. Therefore, further registration parameters, namely the number of resolution levels, were studied in terms of the attainable registration accuracy. Figure 5.3 displays the results obtained for a multistage registration of CBCT to CT (Figure 5.3a), T1 to CT (Figure 5.3b) and T2 to CT (Figure 5.3c) using between 1 and 4 spatial resolution levels and AMMI as similarity measure. As expected, a higher number of resolution levels results in a higher median registration accuracy. However, a registration based on 3 resolution levels results in a similar registration accuracy as the default setting of 4 levels, while simultaneously resulting in a lower computation time. This can be an important aspect considering interventional procedures which require accurate as well as time efficient registration methods. In our experiments, an additional layer increased the computation time for the registration of CBCT to CT by $61.9 \pm 19.3\,\%$, for the registration of T1 to CT by $30.9 \pm 5.9\,\%$ and for the registration of T2 to CT by $27.1 \pm 5.8\,\%$.



(a) CBCT to CT (AMMI, with liver mask)



(b) T1 to CT (AMMI, without mask)



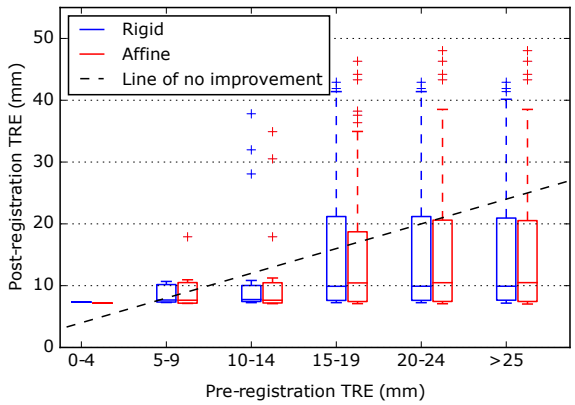
(c) T2 to CT (AMMI, without mask)

Figure 5.3: **Number of Resolution Levels**: Results obtained for the registration of (a) CBCT to CT, (b) T1 MRI to CT and (c) T2 to CT for one of the patients using different numbers of resolution layers for a multistage registration.

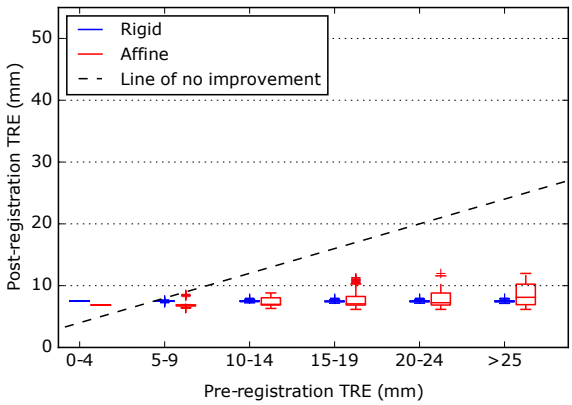
5.1.5 Rigid vs. Affine Registration

As additional evaluation step, the capture range based on the TRE of the manually chosen landmarks used for the point-based ground truth registration was estimated for a rigid registration method and compared to the capture range obtained for an affine method. The results for the registration of CBCT to CT (Figure 5.4a), T1 to CT (Figure 5.4b) and T2 to CT (Figure 5.4c) show that the affine registration performs very similar to a rigid approach and therefore does not provide an advantage for the data used in the evaluation.

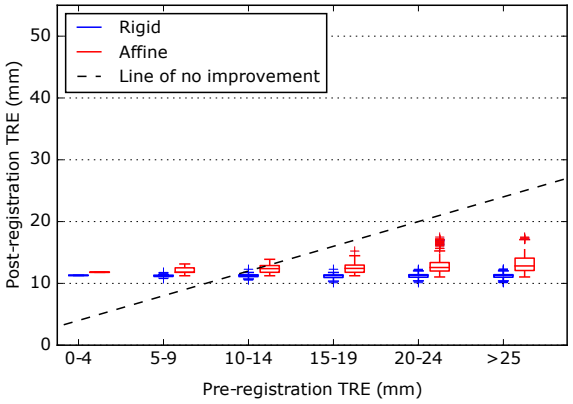
To sum up the findings of the evaluation of different rigid registration methods, the results show that the performance of an initialization before starting a registration process can be essential. The choice of initialization methods depends on the image data that is to be registered: for the registration of CBCT to CT an initialization based on the superposition of liver center (or generally expressed - the center of target structures) results in the best prealignment of the images whereas the superposition of geometric centers yields an optimal prealignment for images covering a same FOV as e.g. MRI and CT data. Concerning the choice a similarity metric, the best results were



(a) CBCT to CT (AMMI, with liver mask)



(b) T1 to CT (AMMI, without mask)



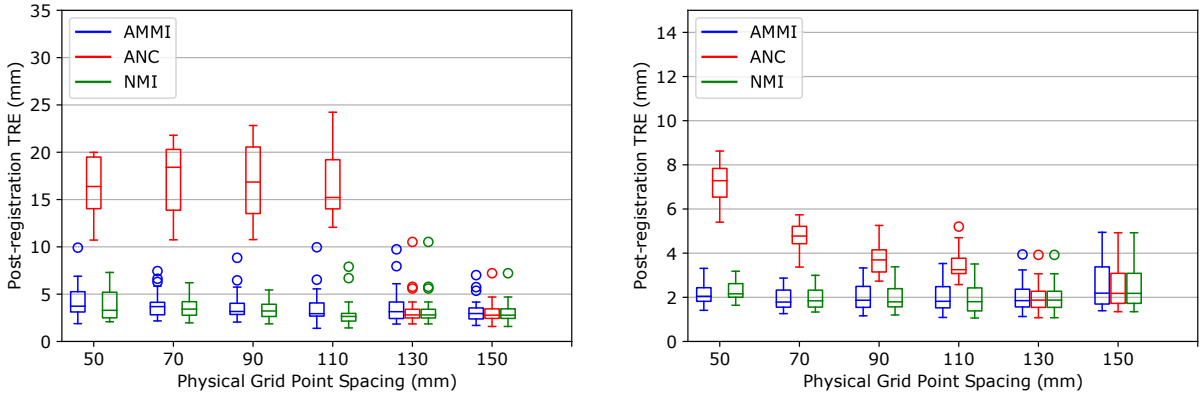
(c) T2 to CT (AMMI, without mask)

Figure 5.4: **Rigid vs. Affine Registration:** Results obtained for the registration of (a) CBCT to CT, (b) T1 MRI to CT and (c) T2 to CT for one of the patients using a rigid and a affine registration relying on the manually chosen landmarks used for the point-based ground truth registration of the data.

obtained using a NMI metric for all modality combinations and patients. Moreover, the results show that three is the optimal number of resolution levels, since it enables an appropriate trade-off between registration accuracy and computation time of the registration method. Especially for the registration of CBCT to CT, the introduction of an image mask during registration was beneficial to achieve a higher registration accuracy.

5.2 Evaluation of Nonlinear Image Registration

The second part of the evaluation methodology focused on non-linear registration methods. The results obtained for the registration of each of the five time steps of the respiratory cycle of the XCAT phantom to all remaining time steps are accumulated and visualized in a single figure. Due to the nature of the displacement in the five respiratory phases, the initial median TRE before registration results ranges either between $5 - 14\,\text{mm}$ or $20 - 29\,\text{mm}$ causing the non-continuous spectrum in the figures shown in this chapter.



(a) CBCT to CT (without mask)

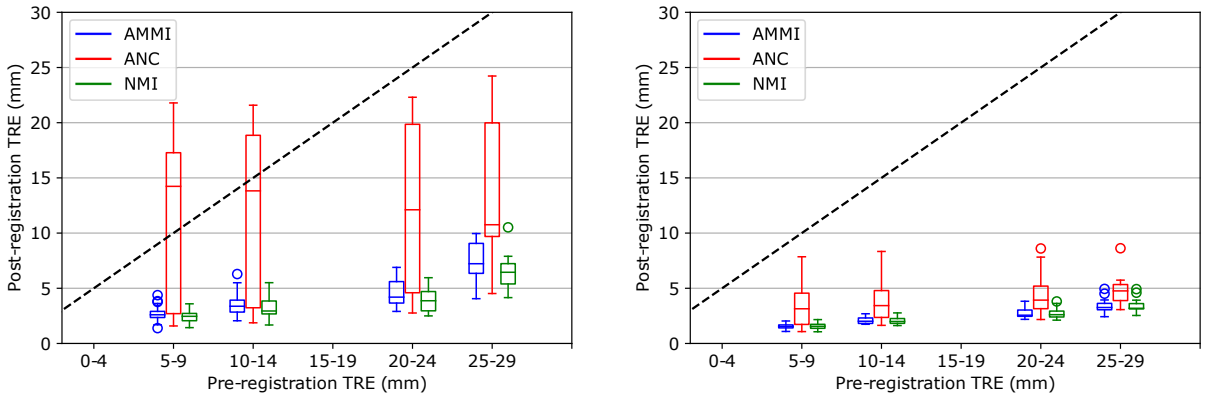
(b) T1 to CT (without mask)

Figure 5.5: **Physical Grid Spacing:** Results obtained for the nonlinear registration of the image data of each point of the respiratory cycle to the data of all other time points by using the three similarity measures AMMI, ANC and NMI as well as a physical grid spacing of the B-spline transform between 50 to 150 mm.

5.2.1 Physical Grid Spacing

Since the physical grid spacing of the applied spline transform generally has a high impact on the achievable registration accuracy, its impact on the registration of abdominal scans has been studied by varying it from 50 to 150 mm in steps of 20. mm. The results obtained for three different registration methods using AMMI, ANC and NMI as similarity measures are shown in Figure 5.5. Concerning the physical grid spacing, the results show that the grid spacing has only a small influence on the registration accuracy for the multimodal registration of abdominal scans. The median TRE for both, the registration of CBCT to CT as well as T1 to CT, is around ~ 2 mm. However, the standard deviation of the TRE increases with an increasing physical grid spacing. In general, a smaller physical grid spacing for the same image volume indicates a higher number of grid points and thus, a higher number of degrees of freedom of the spline transform. This also leads to a higher computation cost that makes the registration time-consuming. It is good practice to choose the grid spacing in correspondence to the size of the geometrical deformations which are expected. In the presented case, the best trade-off between a flexible registration and a reasonable number of grid points is given by a registration using a physical grid spacing of 110 mm. This setting results in a low mTRE of approximately 3 mm for CBCT to CT and approximately 2 mm for T1 to CT after registration. Therefore, this setting yields a high registration accuracy and was used for all following experiments.

The experiments were performed using three different similarity measures AMMI, ANC and NMI. It has to be noted, that the number of grid points highly influences the registration based on ANC, which is considered a monomodal similarity measure and therefore not suitable for multimodal applications. The results show, that for an increased grid spacing (larger than 130 mm), ANC yields similar results as the metrics based on mutual-information. A possible explanation is given by the fact that a larger grid spacing automatically implies less flexibility of the algorithm. If the algorithm has a high flexibility and is able to intensely warp the image, it possibly shifts the grid points to random positions in case an inappropriate similarity measure is used. Since ANC is



(a) CBCT to CT (without mask, grid spacing: 110 mm) (b) T1 to CT (without mask, grid spacing: 110 mm)

Figure 5.6: **Similarity Metrics:** Results obtained for the nonlinear registration of the image data of each point of the respiratory cycle to the data of all other time points by using the three similarity measures AMMI, ANC and NMI.

considered a monomodal similarity measure, this is possibly the case here. A higher spacing and less grid points prevent the algorithm from randomly warping the image which leads to a smaller TRE. The impact of similarity measures on deformable registrations is further discussed in the following section.

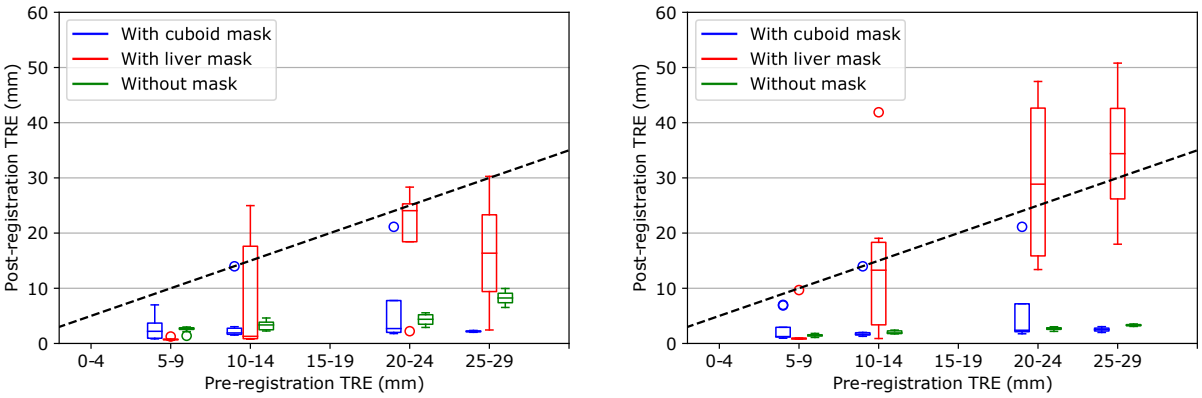
5.2.2 Similarity Metrics

As for the evaluation of linear registration methods, AMMI, ANC and NMI have been investigated as similarity measures for nonlinear registration of abdominal scans. The results have been grouped according to the initial median TRE in steps of 5 mm before the registration and are displayed in Figure 5.6. For both, the registration of CBCT to CT and T1 to CT, the metrics based on mutual information yield a lower median postregistration TRE than ANC, with NMI slightly outperforming AMMI for the registration of CBCT to CT.

The registration of CBCT to CT based on ANC a similarity metric results in a high postregistration of $\sim$ 15 mm, making the metric not suitable for this modality combination. However, the results shown in Figure 5.5 indicate that the registration performance using ANC can potentially be improved by increasing the physical grid spacing of the spline transform. Moreover, ANC yields promising results for the registration of T1 to CT which is consistent with the results obtained for linear registration methods (shown in Section 5.1.2).

5.2.3 Application of Masks

The next part of the evaluation was dedicated to study the impact of a fixed image mask and its shape on the registration accuracy. The results of the nonlinear registration based on AMMI without a mask, with a cuboid fixed image mask that covers the region of the liver and a liver mask are grouped according to the initial TRE and shown in Figure 5.7. It has to be noted the



(a) CBCT to CT (AMMI, grid spacing: 110 mm)

(b) T1 to CT (AMMI, grid spacing: 110 mm)

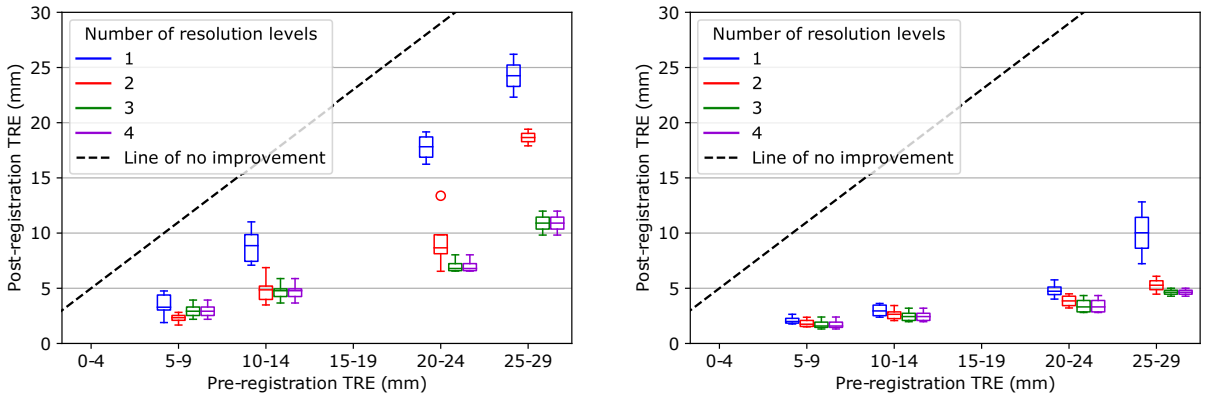
Figure 5.7: **Application of Masks:** Comparison of results obtained for the nonlinear registration of each point of the respiratory cycle to the other time points without mask, using the liver segmentation and using the cuboid segmentation as fixed image mask during the registration.

registration of the synthetic CBCT to the synthetic CT does not require an image mask to prevent the registration from failure, in contrast to the CBCT to CT registration of actual patient data as presented for linear registrations. This is due to the fact, that the synthetic images are intrinsically registered and display a very similar morphology and a lower level of noise which facilitates the registration task.

The results show that using a liver mask drastically decreases the registration accuracy. This can be explained by the fact that by using a liver mask, the similarity metric is only evaluated in the liver region during registration. Since the liver is considered to be a very homogenous organ with only few internal vessel structures and deformable registrations rely on local information to generate an appropriate deformation field, this limitation causes an increase of the postregistration TRE. Thus, using a larger cuboid mask leads to an improvement of the registration accuracy whereas the best accuracy for nonlinear registration is obtained without a fixed image mask. This is valid for the registration of CBCT to CT as well as T1 to CT. The results are consistent with the results obtained for linear registration for T1 to CT, but deviate for the registration of CBCT to CT whose performance for linear registrations could be improved using a liver mask (Sec. 5.1.3).

5.2.4 Number of Resolution Levels

The last parameter that was investigated for nonlinear registrations is the number of resolution levels in a multistage registration. As for linear registrations, it is important to chose the number of resolution levels as a trade-off between registration accuracy and registration speed. The registration results for resolution levels between 1 up to 4 for the nonlinear registration of CBCT to CT and T1 to CT are displayed in Figure 5.8. As for linear registration of abdominal scans, the registration accuracy generally decreases with a decreasing number of resolution levels. However, the multistage registration based on 3 resolution layers yields the same registration accuracy as the registration using 4 layers and thus, represents the best setting for the nonlinear registration of abdominal scans.



(a) CBCT to CT (AMMI, no mask, grid spacing: 110 mm) (b) T1 to CT (AMMI, no mask, grid spacing: 110 mm)

Figure 5.8: **Number of Resolution Levels:** Results obtained for the nonlinear registration of each point of the respiratory cycle to the other time points using different numbers of resolution layers for a multistage registration.

In summary, the results obtained for the evaluation of non-linear registration methods are in good agreement with the results obtained for the evaluation of linear methods. As for linear registration approaches, the highest registration accuracy was obtained using NMI as similarity measure and three layers of resolution. Moreover, the results show that the physical grid spacing of the spline transform only has a small impact on the registration accuracy and that a spacing of 110 mm is appropriate for the registration of abdominal data. However, in contrast to the results obtained for rigid registration methods, the introduction of an image mask decreased the registration accuracy.

5.3 Novel Similarity Metrics

Since previous results of the registration evaluation show that the choice of the similarity measure has a significant impact on the capture range of the registration method, two alternative similarity measure are proposed within the frame of this thesis. Both metrics are characterized for different types of geometric transformation using multimodal data pairs. The results are presented in the following.

5.3.1 Similarity Metric based on *Histograms of Oriented Gradients*

The evaluation of the HOG-based similarity metric is divided into two parts: the first part aims to investigate the behavior of the isolated similarity measure by sampling the parametric space for translation and rotation transformations. In the second part of the evaluation, the metric is incorporated in a complete registration algorithm and its performance for the rigid registration of abdominal CBCT to CT scans is studied in terms of the estimation of the capture range of the method.

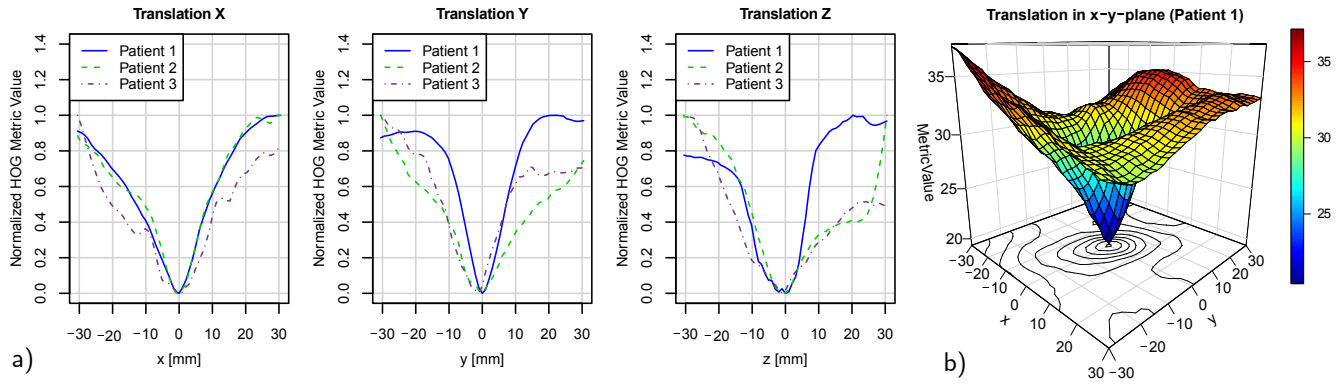


Figure 5.9: a) Parametric space obtained with the HOG-based similarity metric for the translation in x-, y- and z-direction for 3D CBCT/CT data of each patient. The metric values were projected in a range from $0-1$ to enable a comparison between the parametric spaces. b) Parametric space obtained with the HOG-based algorithm for the translation in the x-y-plane for one data set.

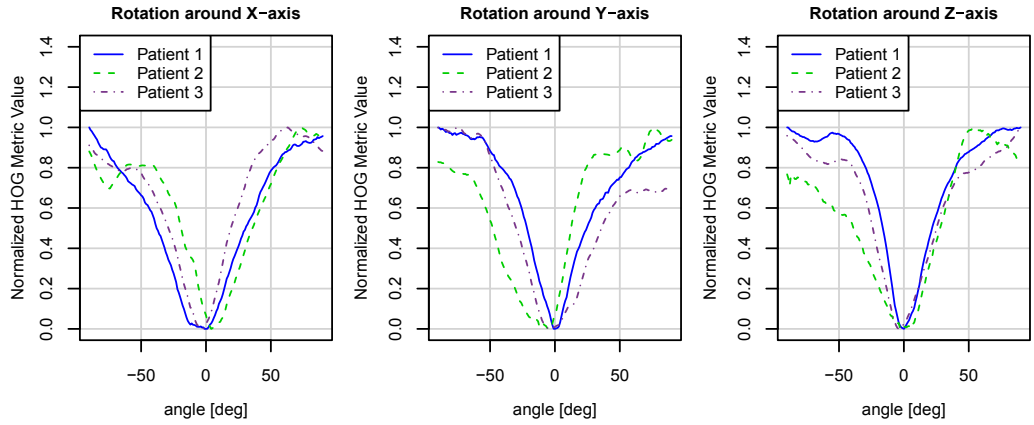


Figure 5.10: Parametric space obtained with the HOG-based similarity metric for the rotation around the x-, y- and z-axis for 3D CBCT/CT data of each patient.

Parametric Space

The parametric space plots obtained for the translation of the 3D-CBCT (set as source image) in x-, y- and z- direction relative to the 3D-CT (set as fixed image) of each data pair for the HOG based similarity metric are displayed in Figure 5.9 a). The metric values were normalized and projected in a range from $0-1$ in order to enable a comparison of the parametric spaces shapes obtained for each pair of patient data sets. For all data sets, a distinct minimum can be identified that indicates the position of the gold standard determined by the preregistration. The HOG-based metric results in smooth parametric cone surfaces for all three pairs of data which display only few notable local minima for translations up to $\pm 10\,\mathrm{mm}$. However, local minima are identifiable for larger translations and the general shape of the parametric hyper-cone surface (exemplary shown in 5.9 b) for translations in x- and y-direction) depends on the image data used for the evaluation. A similar metric behavior is observed for the case of rotational transformations for which the results are shown in Figure 5.10. For rotations up to $\pm 50\,^{\circ}$, the metric values in parametric space are essentially smooth and monotonic, but show a few notable local minima for larger rotations.

To evaluate the performance of the HOG-based metric in comparison to a well-established similarity

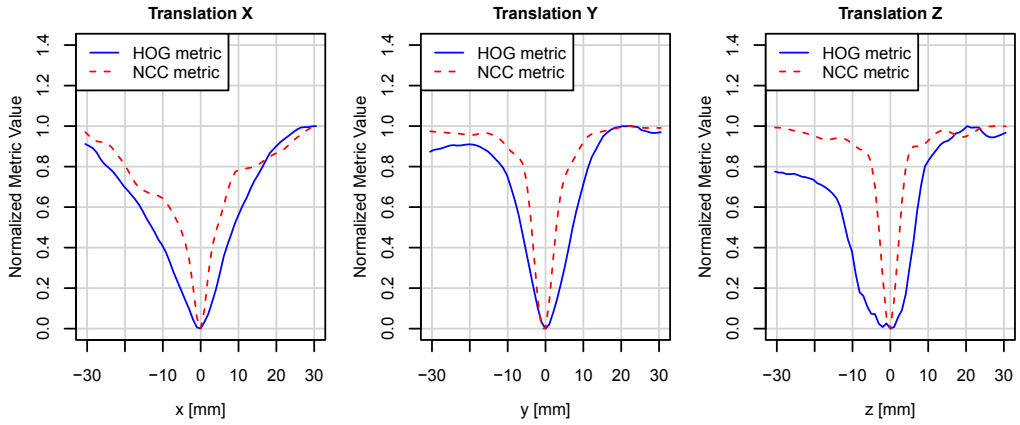


Figure 5.11: Parametric space obtained with the HOG-based similarity metric and the NCC metric for the translation of the moving image in x-, y- and z-direction for one CBCT/CT data pair. As in Figure 5.9, the metric values were projected in a range from $0-1$ to enable a comparison between the two metrics. The plots show that the HOG metric results in a broader parametric cone than the NCC metric, making it easier to find during the optimization process.

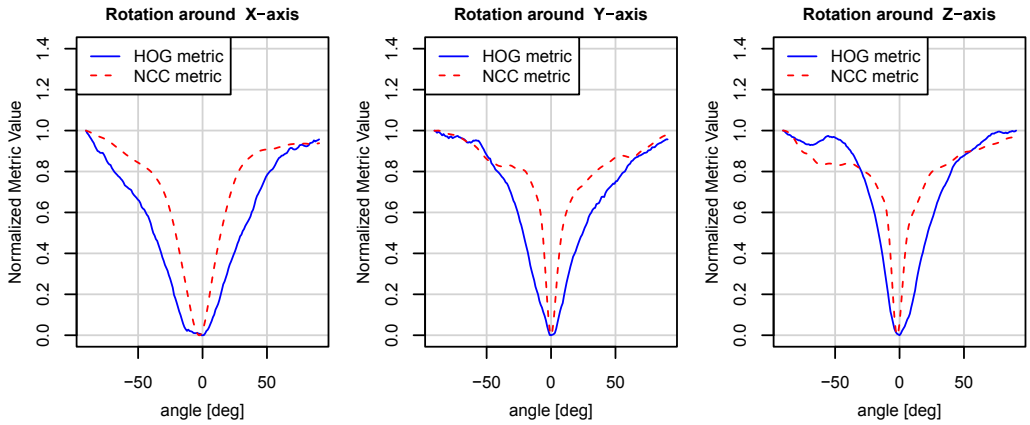


Figure 5.12: Parametric space obtained with the HOG-based similarity metric and the NCC metric for the rotation of the moving image around the x-, y- and z-axis for one of the CBCT/CT data pairs. Again, the metric values were projected in a range from $0-1$. As observed for the evaluation for translation transformations, the plots show that the HOG metric results in a broader parametric cone than the NCC metric.

metric, the normalized parametric space plots obtained for translation and rotation transformations using the HOG as well as the NCC metric for a single data set are shown in Figure 5.11 and figure 5.12, respectively. Both methods result in smooth parametric cone surfaces for both types of transformations with a distinct minimum. In general, the parametric cone obtained for the HOG metric is broader than the one obtained for the NCC metric, possibly facilitating the identification of the global minimum during the optimization process.

Capture Range

Next, the registration accuracy of the method is investigated by registering the 3D-CT and the 3D-CBCT data pairs of all three patients after artificial displacements relative to a reference position. The registration capture range is estimated for translation and rotation transformations obtained for the registration relying on the HOG-based similarity metric and the results are shown in Figure

5.13 a) and 5.14 a), respectively.

For all data pairs, it can be stated that the registration using the HOG-based metric yields an average TRE below the successful registration threshold (set to a median TRE of 3 mm) for initial mean displacements (initial mTRE before registration) of up to 10 mm for translation and rotation transformations. The results obtained for larger initial displacements show that the final mTRE increases for increasing start mTREs, but that the HOG-based registration still improves the alignment of the 3D-CT to the 3D-CBCT for all patients for both, the translational and rotational case. However, the standard deviation of the final mTRE also increases for increasing initial displacements which is additionally demonstrated by the percentage of successful registrations per displacements. These results are shown in Figure 5.13 b) for the case of translation and in Figure 5.14 b) for rotation transformations. The mean rate of successful registrations constantly decreases for increasing displacements as it gets more complicated for the optimizer to reach the global minimum of the metric. This can be related to the fact that the amount of local minima in the parametric space in which the optimizer can possibly get caught increases for larger displacements, as discussed in the previous section 5.3.1.

Figure 5.15 a) and figure 5.16 a) show the results of the translation as well as the rotation capture range obtained for the evaluation using the HOG- and NCC-based registration methods. Similar to the HOG-based metric, the NCC-based algorithm yields an average final mTRE below the threshold of successful registration for small initial displacements. Concerning translation transformations, the final mTRE obtained with the HOG metric is still below this threshold for displacements in the range between $10-15$ mm, whereas the final mTREs obtained with NCC start to increase. For larger displacements, the final mTRE increases approximately linear for both algorithms. This behavior can also be observed when examining the percentage of successful registrations displayed in Figure 5.15 b). Although both algorithms yield similar success rates for small initial displacements, the HOG-based metric outperforms the NCC-based algorithm for medium displacements in terms of the percentage of successful registrations. A reversed metric behavior is observed for the evaluation of the rotational capture range shown in Figure 5.16 b). Again, both algorithms yield similar registration results for small displacements, but the NCC-based algorithm offers a higher probability for a successful registration than the HOG-metric for medium displacements in the range between $10 - 20$ mm. Moreover, figure 5.15 a) and figure 5.16 a) show that the standard deviation for the NCC-based registration is constantly smaller than the standard deviation obtained for the HOG-based registration, indicating a smaller probability of large misregistrations.

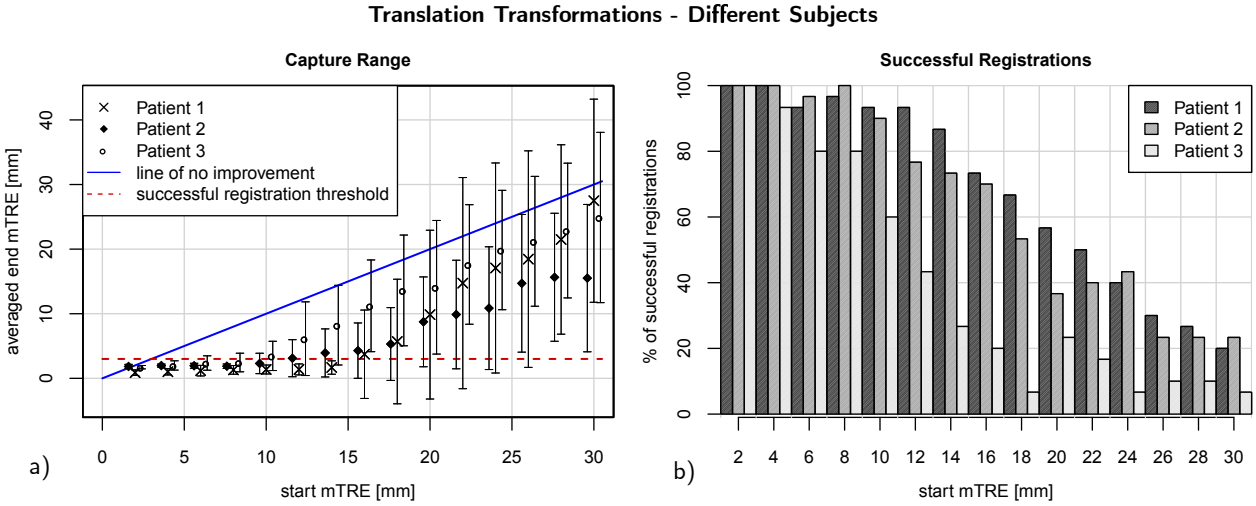


Figure 5.13: **Translation Transformations:** a) Capture range using the HOG-based similarity metric for translation transformations for the data of three patients. Although the values for the start mTRE were the same for every data pair, the results were plotted separately to avoid an overlapping of the errorbars. b) The barplot displays the percentage of successful registrations achieved for each data pair. The threshold for a successful registration was set to 3 mm.

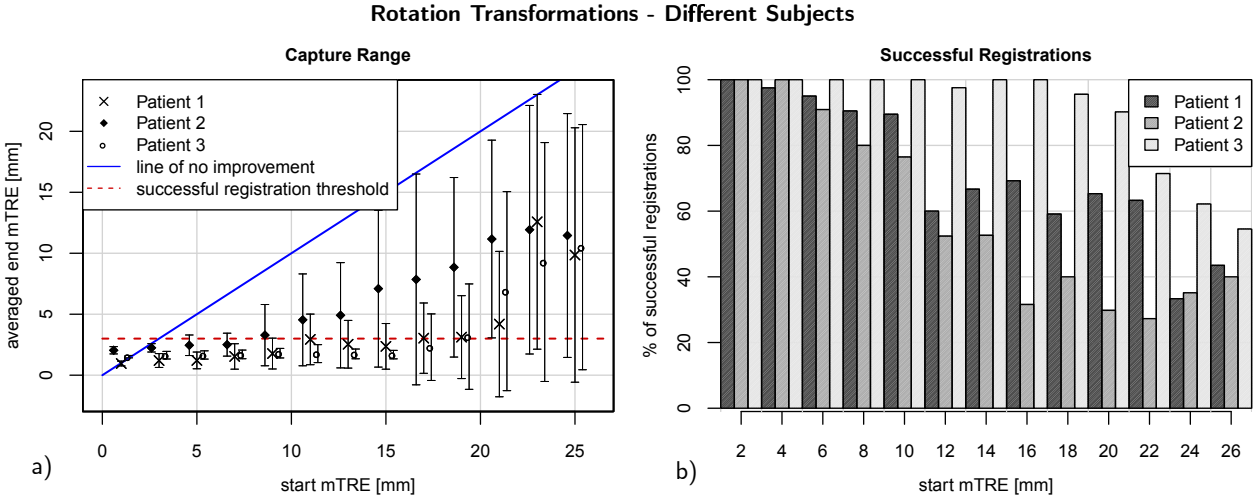


Figure 5.14: **Rotation Transformations:** a) Capture range using the HOG-based similarity metric for rotation transformations. for the data of three patient. b) The barplot displays the percentage of successful registrations achieved for each data pair. The threshold for a successful registration was set to 3 mm.

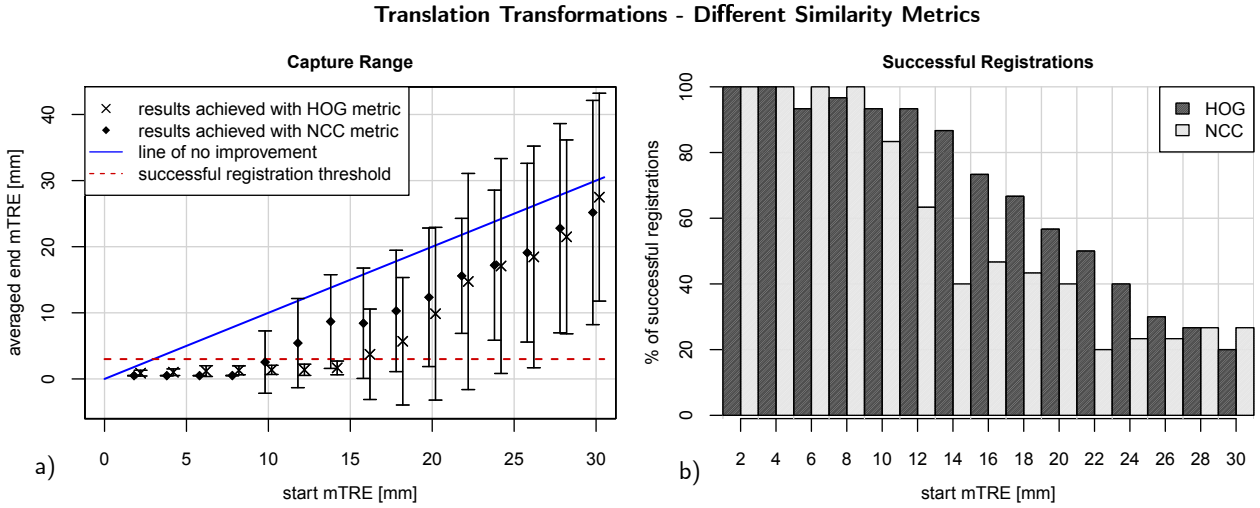


Figure 5.15: **Translation Transformations:** a) Capture ranges using the HOG-based similarity metric and NCC metric for translation transformations. b) The barplot displays the percentage of successful registrations achieved with both metric types. The threshold for a successful registration was set to 3 mm.

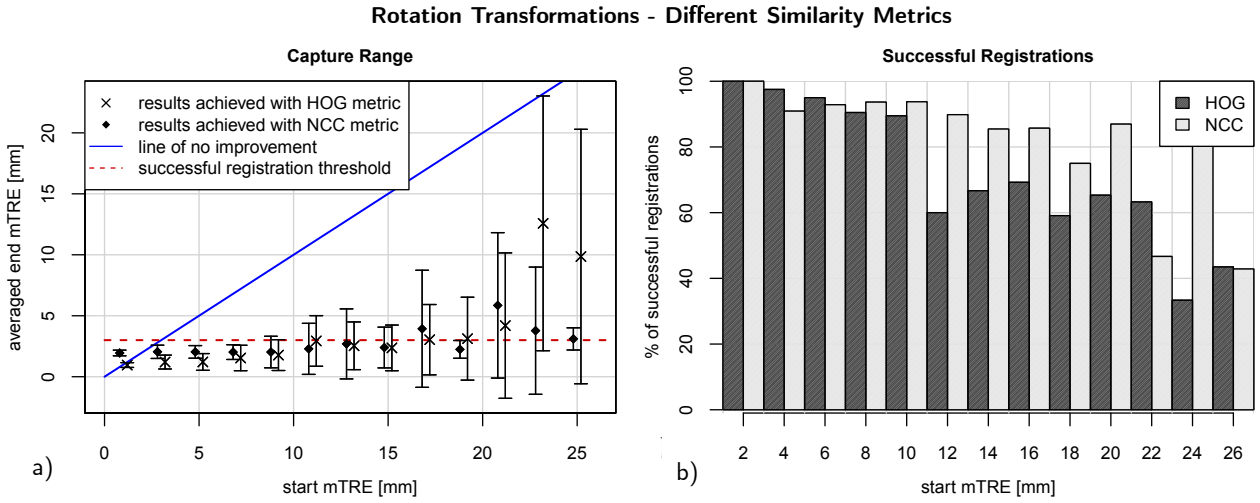


Figure 5.16: **Rotation Transformations:** a) Capture ranges using the HOG-based similarity metric and NCC metric for rotation transformations. b) The barplot displays the percentage of successful registrations achieved with both metric types. The threshold for a successful registration was set to 3 mm.

However, since the comparison of both similarity metrics relies on the registration of a very limited number of 3D CT and CBCT data pairs, no general conclusion indicating which metric is more suitable for the registration of this modality combination can be drawn from the results. Nevertheless, the results show that the HOG-based similarity metric is generally suitable to be used as similarity metric for the registration of abdominal CBCT to CT data. Moreover, the registration results obtained with the HOG-based registration method indicate a performance similar to conventional similarity measures such as NCC.

A major drawback of this novel similarity metric is its high computation time which is mainly caused by the time required for the extraction of the 3D-HOG descriptor ranging between 10 to 20 s for a conventional CT depending on the exact size of the input image. But due to the cell-based structure of the HOG algorithm, it is a suitable candidate for an optimized parallel implementation on a GPU. A GPU-based computation of the extraction of a 3D-HOG descriptor has been realized in cooperation with colleagues from the Department of Computer-Assisted Clinical Medicine, Heidelberg University. This could speed-up the computation of the 3D-HOG descriptor to approximately 1 s for abdominal volumes with 65 mio voxel. The results have been published in the proceedings of the *3rd Conference on Image-Guided Interventions & Fokus Neuroradiologie (IGIC 2017)* [195]. However, since a registration process corresponds to an iterative optimization procedure, the 3D HOG descriptor has to be computed multiple times per registration leading to an unbalanced trade-off between computation time and registration accuracy of the metric. Therefore, the applicability of this novel metric in a clinical context is rather limited and no further experiments on multimodal image data were performed.

5.3.2 Similarity Metric based on a Siamese Neural Network

The results presented in this chapter aim to further understand and characterize a learned similarity measure based on a siamese network architecture for different modality combinations and morphological data. The performance of the Siamese Deep Metric (SDM) is evaluated using modality combinations of three different data types:

- MRI brain scans (T1/T2),
- multimodal synthetic image data of the abdomen (CBCT/CT, T1/CT),
- and (pre-)interventional multimodal patient data of the abdomen (CBCT/CT).

The following chapter is divided according to these three types of evaluation data. The first two data types actually served as training data for the siamese network to estimate similarity between T1/T2-MRI brain scans, as well as synthetic CBCT/CT and T1/CT abdominal scans. To further investigate the transferability of the metric, the performance for a model trained with synthetic data is evaluated on real patient CBCT/CT data. For each modality combination, the parametric space for translation and rotation transformations has been sampled for the different types of metric evaluation presented in Figure 4.9.

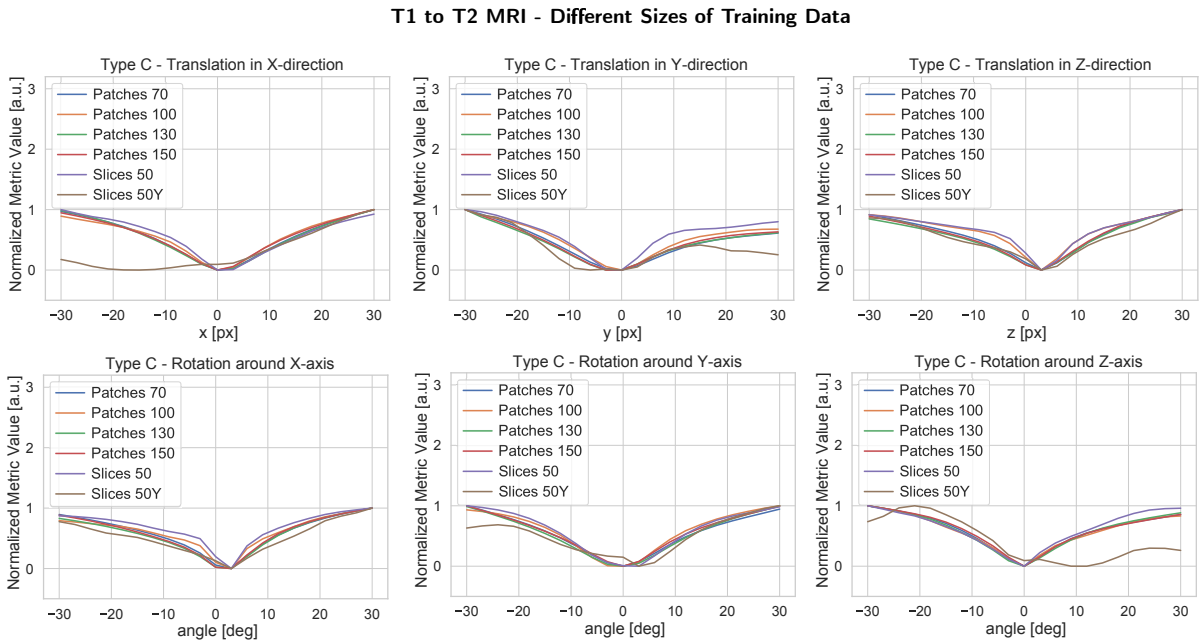


Figure 5.17: Parametric space obtained for a pair of T1 and T2 MRI data using the SDM trained with different patch sizes, for translation and rotation transformations.

To investigate the influence of the training data on the performance of the SDM, the siamese network has been trained with image pairs of different sizes as listed in Table 4.5. Due to reasons of clarity, the influence of the patch size of the training data on the performance of the similarity measure is not shown for all metric evaluation types. Type A and B (see figure 4.9) are mostly relevant for the similarity estimation between image volumes that cover the image regions of very similar physical size. However, this is not often the case for medical image data acquired during clinical routine. The difference in physical space covered by an imaging modality is especially pronounced for pre-interventional data and intra-interventional CBCT, since the CBCT is often characterized by a FOV that is limited to the anatomical region of interest. So the evaluation type C represents the most relevant metric evaluation scenario in a clinical context with regards to the use-case of interventional image registration. Hence, the results for all data types that will be shown for the performance study of SDM models trained with different patch sizes are limited to type C for all modality combinations. If not mentioned otherwise, the results are also valid for evaluation types A and B.

The results obtained for the novel Siamese metric are compared to the results of traditional multimodal similarity measures, corresponding to a Normalized Gradient Field metric (NGF) and the Advanced Mattes Mutual Information (AMMI) with and without consideration of an image mask during metric evaluation. For all results, the numerical values obtained for the similarity metrics were projected in a range from 0 to 1 to enable a comparison between the different metrics.

T1/T2-MRI Brain Data

As discussed in Section 4.2.4, the parameter sampling in an ideal case results in a conical-shaped parametric space landscape with a distinct extremum, either minimum or maximum depending on

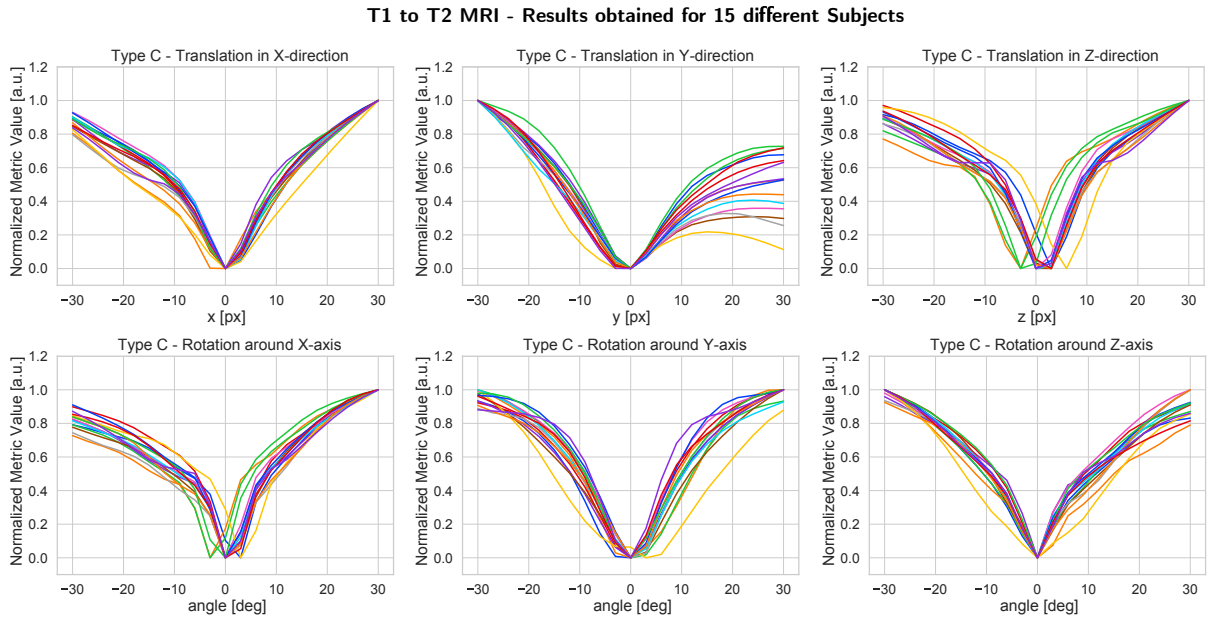


Figure 5.18: Parametric space obtained for the evaluation of different T1 and T2 MRI data pairs using the SDM trained with a patch size of $100 \times 100 \times 100$ pixel for translation and rotation transformations. Each color corresponds to the parametric space obtained for one image pair.

the employed similarity measure, and a smooth surface without local minima which potentially hinder the optimization process.

Figure 5.17 displays the results obtained for the evaluation of Siamese metric models trained with different sizes of input data. The data varied in terms of image dimension of the input pairs ranging from cubic patches (denoted as *Patches* in the figure legend) to 3D slices (denoted as *Slices* in the figure legend). A complete list of the input dimensions is given in Table 4.5. The main difference of these patches is the fact that they contain different degrees of image information. Since the different input patches are sampled from the same native data set (as explained in Section 4.3.2), cubic patches with a size of $70 \times 70 \times 70$ pixel contain less structural information than e.g. patches with a size of $130 \times 130 \times 130$ pixel. The results show, that all SDM models trained with different input sizes yield a very similar performance with the exception of the model the trained with 3D slices *Slices 50 Y*. These slices are characterized by a size $50 \times 238 \times 188$ whereby the slices were extracted along the sagittal axis of the brain. The difference between these slices and the other training data is the fact, that the T1 and T2 MRI do not cover the same field of view in axial direction for all patients. In some cases, the T1 MRI covers the entire head region from skull to neck whereas the corresponding T2 MRI only covers the skull. Although all training patches are exclusively extracted above the neck region, some T2 *Slices 50 Y* contain a small image region which is characterized by zero values due to the image resampling during pre-registration of the data. These non-valid image regions are contained in the T1 *Slices 50 Y* for several data pairs which potentially hinders the similarity learning for these cases. Although such image regions may also appear in patches with other sizes, the ratio of patches containing such regions to patches without these regions is possibly higher for *Slices 50 Y*, since the extraction possibilities of these patch pairs in the original data sets are limited due to their size. As a result, the parametric cone for these cases is not as smooth as for the other patch sizes.In general, the results indicate that

the size of the training data and thus, the contained image information has only a small impact on the similarity learning, in case the training patches are only extracted in valid image regions of the native data set. Moreover, the results suggest that the number of training pairs does not affect the performance of the SDM, since the extraction of the different sized training data from the native MRI scans resulted in a varying number of training pairs as listed in Table 4.5 according to the size of the extracted training data. Due to the performance similarity of all SDM models trained with different input sizes, the results shown in the following are restricted to the parametric spaces obtained using training data with a medium size of $100 \times 100 \times 100$ pixel.

To ensure that the SDM yields an accurate performance for the similarity estimation between more than one pair of T1/T2 MRI scans, the performance has been evaluated for a total of 15 brain data pairs. The results are shown in Figure 5.18. Each color corresponds to the parametric space obtained for one data pair. The SDM yields a smooth parametric cone for all T1/T2 pairs with a distinct minimum. Although the data pairs were pre-registered as explained in Section 4.2.1, some of the data pairs still display slight misalignments which can be seen by small shifts of the mimima for the translation in z-direction. However, the results indicate that the SDM accurately estimates the similarity between multiple T1/T2 MRI scans for translation and rotation transformations suggesting a general validity of the metric for different testing subjects.

To further analyze the behavior of the metric for both of these transformation types, the parametric cones obtained for one data pair using the SDM model trained with *Patches 100* for all three metric evaluation types A, B and C are shown in Figure 5.19 and Figure 5.20, respectively. The columns of these figures represent the type of metric evaluation A, B or C, whereas the lines show the results obtained for one particular translation direction or rotation axis.

For both transformation types, the Siamese Deep Metric (SDM) results in smooth parametric cones that do not display any local extrema. Moreover, the parametric cones present a significant global minimum that indicates the position of optimal image alignment according to the metric. For all translation and rotation cases, the position of this optimum corresponds to the same position which is identified by the traditional similarity measures such as the NGF metric, as well as AMMI with and without mask. Comparing the shapes of the parametric spaces obtained with different similarity measures, NGF results in a very narrow cone with several discontinuities e.g. seen for the translation in y-direction or along (1,1,0) axis (see Fig. 5.19) as well as the rotation around x- and y-axis (see Fig.5.20). Since this can lead to inaccurate image alignment during the registration process, NGF seems inappropriate for the registration of the T1 to T2 MRI brain scans.

The benefit of an image mask for the metric evaluation is demonstrated by the results obtained for the AMMI metric with and without mask for metric evaluation type B. For this type, the optimal image alignment does not correspond to the position with a maximal image overlap. AMMI without the use of an image mask results in a displacement of the correct minimum of the parametric cone by up to 10, mm as seen for the translation along the y- and diagonal (1,1,0)-axis as well as all rotation transformations. This displacement of the minimum potentially leads to an incorrect transformation estimation during a registration process. Therefore, the results shows that the application of an image mask during for the similarity metric estimation during a registration process can be beneficial to increase registration accuracy.

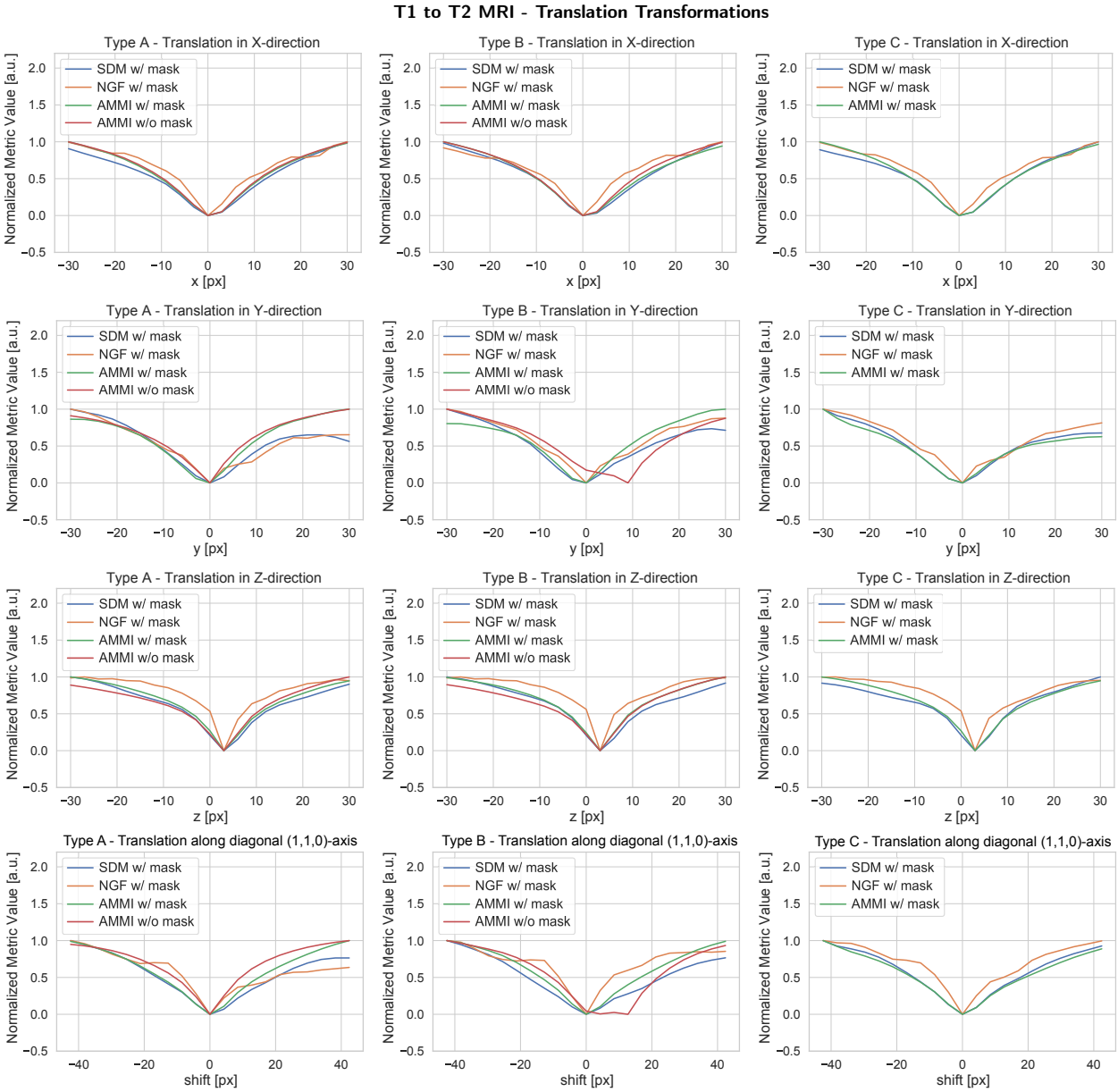


Figure 5.19: Parametric space obtained for a pair of T1 and T2 MRI data using the learned SDM, a NGF metric and a traditional AMMI metric with and without the use of an image mask for translation in x-,y-,z- and diagonal (1,1,0)-direction. The metric values were projected in a range from 0 - 1 to enable a comparison between the different metrics.

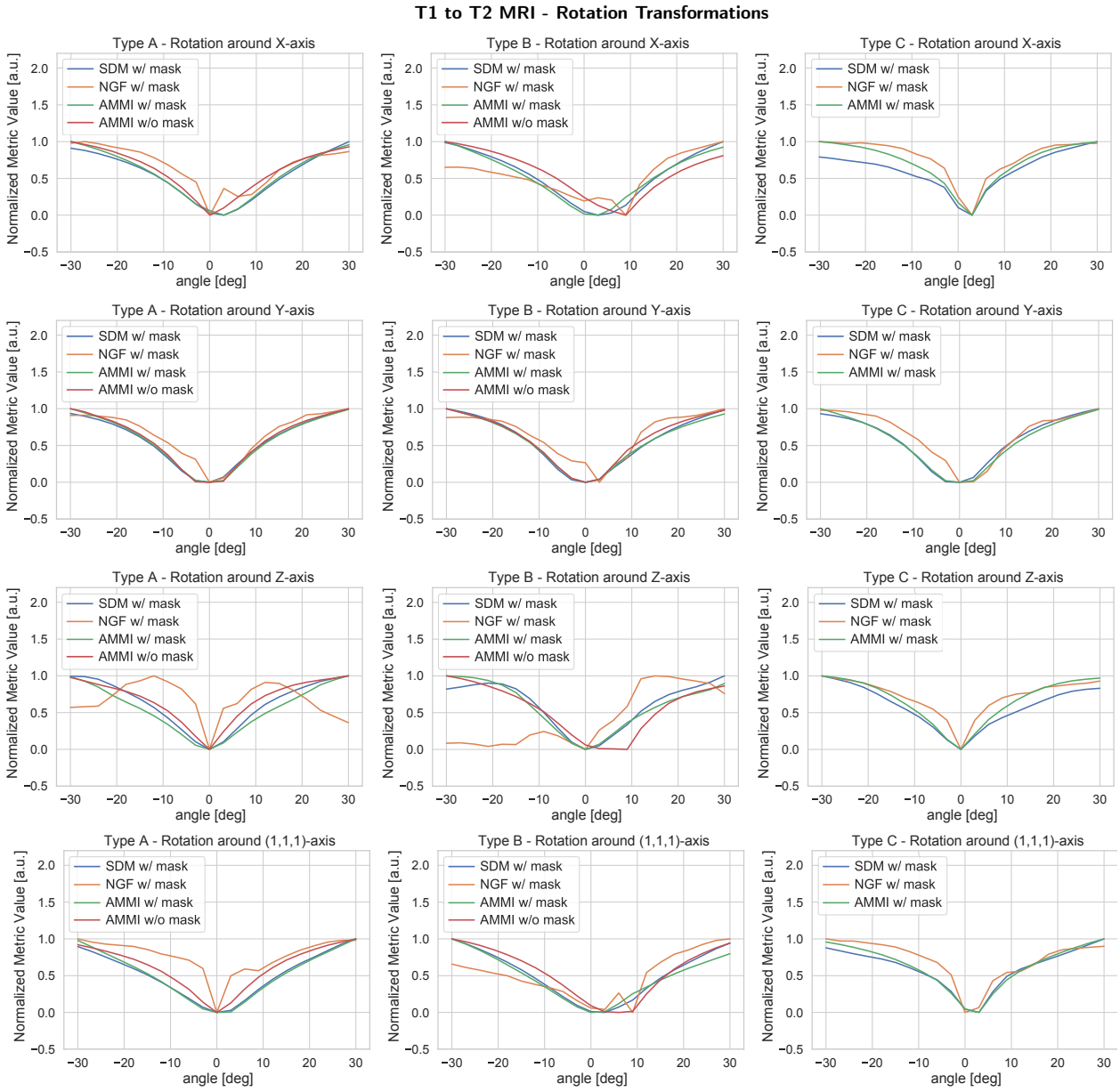


Figure 5.20: Parametric space obtained for a pair of T1 and T2 MRI data using the learned SDM, a NGF metric and a traditional AMMI metric with and without the use of an image mask for rotation around the x-,y-,z- and diagonal (1,1,1)-axis.

Synthetic Abdominal Data

To investigate the behavior of the metric on another type of patient morphology, the SDM has been additionally trained and evaluated for the similarity estimation of multimodal abdominal data. Two different modality combinations have been studied, corresponding to synthetic CBCT/CT and synthetic T1 MRI/CT. For the sake of clarity, only the results obtained for the modality combination CBCT/CT are shown here, whereas the results obtained for T1 MRI/CT can be found in Appendix 7.

As for the MRI brain data, the impact of the training data on the metric performance is evaluated for different input sizes. Figure 5.21 displays the results obtained for the similarity estimation between synthetic CBCT and CT data. In contrast to the results obtained for the brain data, the size of the training patches affects the performance of the metric especially for translation transformations. Here, an increasing size of the cubic patches, corresponding to the *Patches 70* up to *Patches 150*, lead to a broader parametric cone. In the context of image registration, a similarity metric that is characterized by a broad parametric cone can decrease the computation time of the registration method. In such a case, the optimization is potentially accelerated due to the continuous slope in the parametric space indicating the optimization direction for the transform parameters. However, the models trained with large cubic patches, yield no clearly identifiable optimum anymore as seen e.g. for the model based *Patches 150* for translation in z-direction. This could potentially hinder the registration process. To establish an appropriate trade-off between a broad parametric cone as well as a significant optimum, the optimal patch size for the modality combination CBCT/CT is given by a medium patch size. Therefore, the following results are shown for the SDM model trained with a patch size of $100 \times 100 \times 100$ pixel. In contrast to the findings for CBCT/CT, the results obtained for the performance evaluation of SDM models trained with different patch sizes of synthetic T1 MRI/CT (shown in Appendix 7) do not display a dependence from the size of the training data. This is in agreement to the results obtained for brain MRI presented in the previous section. A possible explanation that the training data size only affects the similarity estimation between CBCT/CT is given by the differences in the FOV between these two modalities. As discussed in Section 2.2.1, the CBCT volume is characterized by a ring-shaped FOV in contrast to a CT volume. To take this into consideration during the similarity estimation, the training patches were extracted at positions that not only include regions that contain valid image information, but also partially include this ring artefact. Since all patches are extracted from the same native CBCT/CT volumes, the extraction of larger patches automatically results in more patches that contain segments of this ring artefact. This could possibly hinder the similarity learning and smaller patches that only contain valid image information could be favorable.

The evaluation is performed for multiple subjects. However, the data used for the evaluation of these SDM models corresponds to the synthetic image triplets of five positions of a respiratory cycle that is simulated using the XCAT phantom (the entire generation process is described in Section 4.2.1). Since the morphological changes are only minimal for the different positions of the respiratory cycle including e.g. movements of the diaphragm, the evaluation results are only shown for the data of one image pair instead of all five image pairs. Differences between the results for different data pairs are only marginal as shown in Appendix 7 for the modality combination

Figure 5.21: Parametric space obtained for a pair of synthetic CBCT and CT data using the SDM trained with different patch sizes, for translation and rotation transformations.

CBCT/CT.

The parametric cones obtained for translation and rotation transformations for the metric evaluation types A, B and C based on the model trained with cubic image pairs with a size of $100 \times 100 \times 100$ px are shown in Figure 5.22 and Figure 5.23, respectively. As for the MRI brain data, the parametric cones obtained for both transformation types display no local extrema and show a clearly identifiable minimum indicating the position of optimal image alignment which is in correspondence with the position obtained using traditional similarity measures. Moreover, in agreement to the findings obtained in the previous chapter, the NGF metric results in a very narrow parametric cone which extends into plateaus, as seen e.g. for the translation along the (1,1,0)-axis. The absence of a slope can complicate the optimization process during an image registration, since it leads to a vanishing gradient that complicates e.g. a gradient-descend optimization.

Compared to the results obtained for the MRI data of the brain, the parametric cones obtained for the SDM models trained on abdominal data provide an even smoother surface. A possible explanation is given by the reduced amount of structural information in the abdominal data sets. Images of the brain tissue are characterized by a lot of texture and sudden intensity changes, which increase the complexity of a similarity estimation. However, the results show that the siamese network is able to learn an appropriate similarity estimation for both types of morphology.

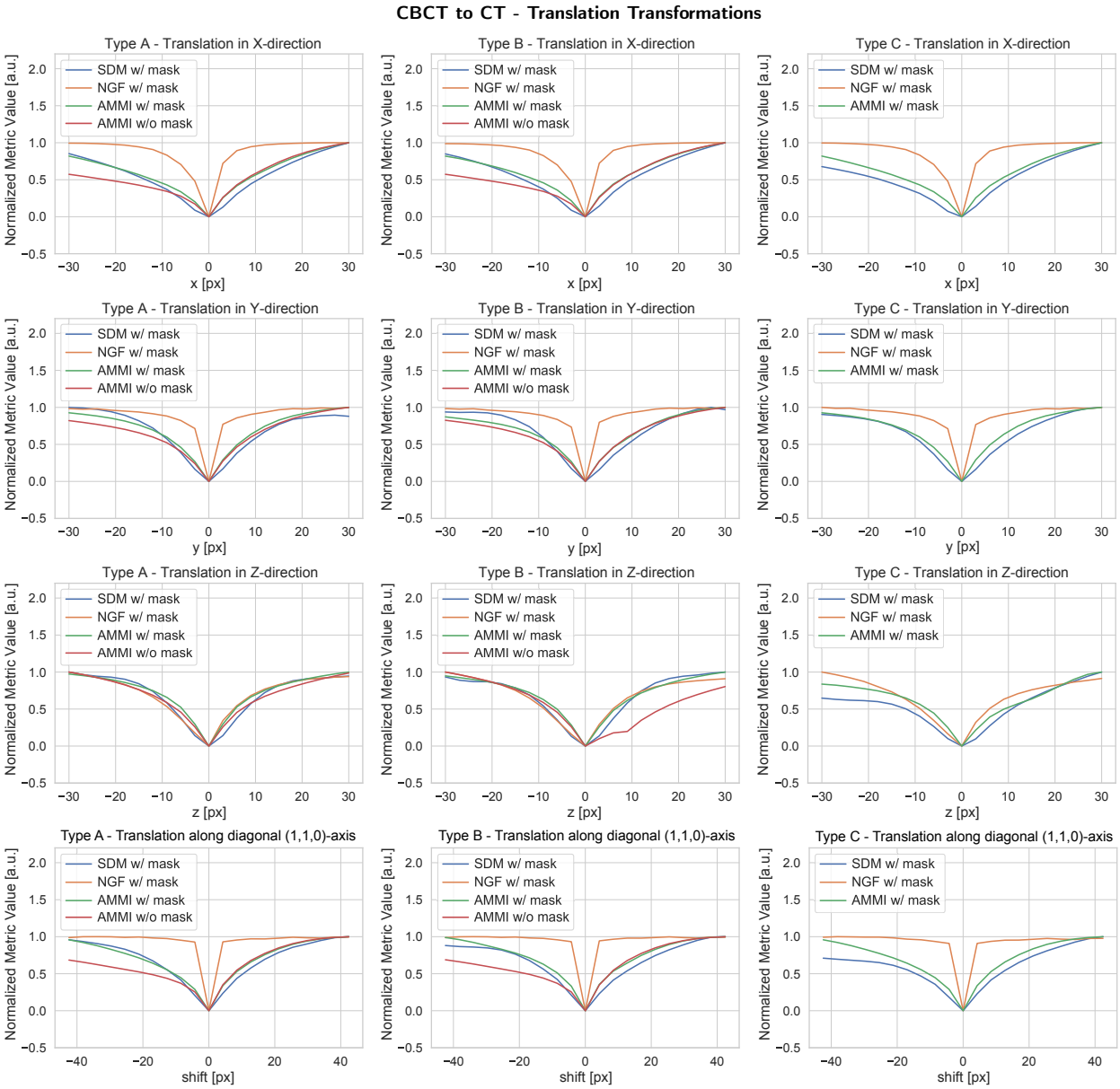


Figure 5.22: Parametric space obtained for a pair of synthetic CBCT and CT data using the learned SDM, a NGF metric and a traditional AMMI metric with and without the use of an image mask for translation in x-,y-,z- and diagonal (1,1,0)-direction. The metric values were projected in a range from 0 - 1 to enable a comparison between the different metrics.

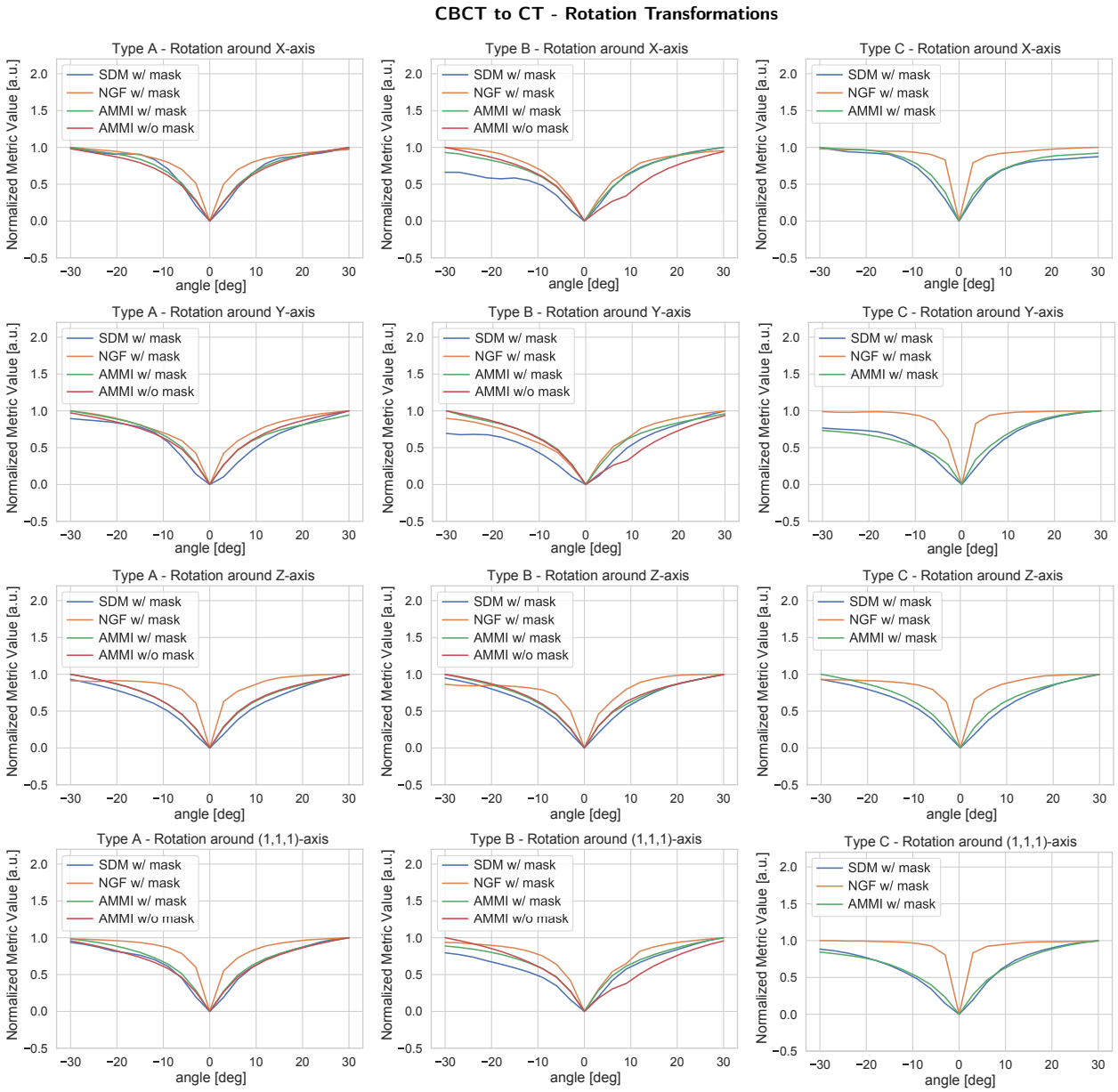


Figure 5.23: Parametric space obtained for a pair of synthetic CBCT and CT data using the learned SDM, a NGF metric and a traditional AMMI metric with and without the use of an image mask for rotation around the x-,y-,z- and diagonal (1,1,1)-axis.

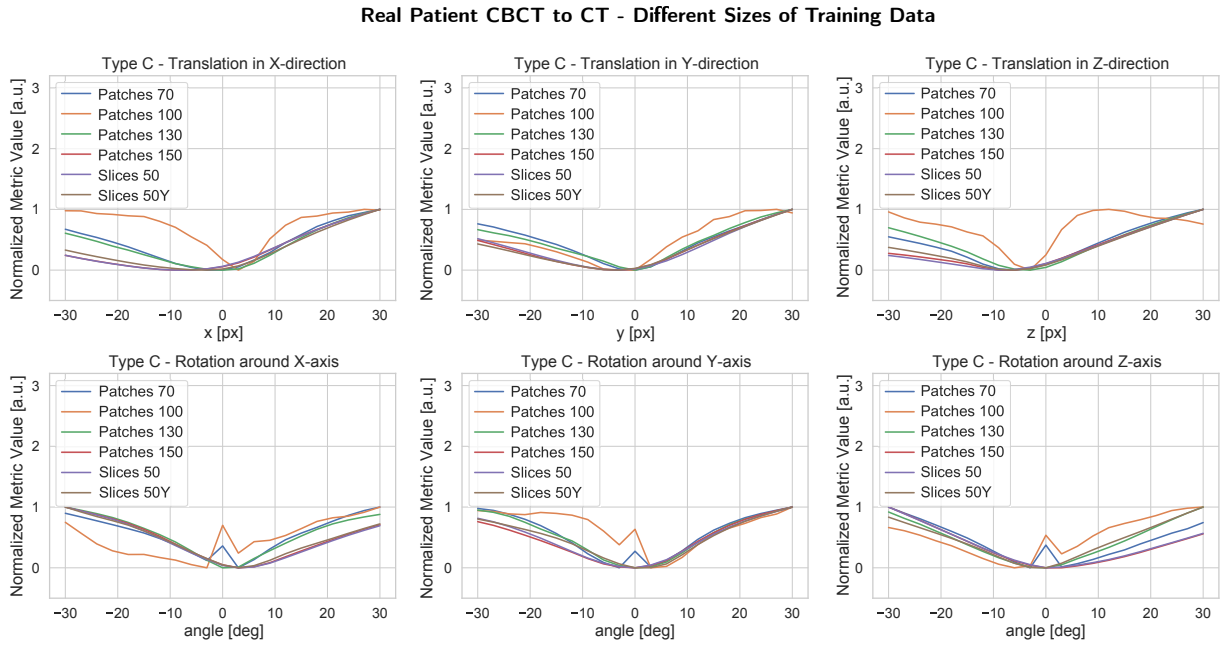


Figure 5.24: Parametric space obtained for a pair of real patient CBCT and CT data using the SDM trained with different patch sizes, for translation and rotation transformations.

Patient CBCT/CT Abdominal Data

At last, the transferability of the SDM models trained with synthetic CBCT/CT-pairs is studied by evaluating the performance for the image similarity estimation of these models on real patient CBCT/CT data.

First, the SDM models trained with different input sizes were evaluated on a single real CBCT/CT pair and the results are shown in Figure 5.24. The results show that the SDM metric generally yields a parabolic parametric landscape for all translation directions and rotation axis whereby the minimum indicates the position of optimal image alignment according to the SDM. This implies a general transferability of the models to real patient data, since the SDM models are able to estimate an optimum. However, in contrast to the results obtained for the evaluation on synthetic CBCT/CT data, there exists a significant variation in the performance of the different SDM models. Especially the models trained with small patch sizes, such as *Patches 70* and *Patches 100* display local discontinuities and not always a significant optimum. This could possibly be related to the fact, that real patient data displays more structure than the synthetic data sets. Taking into consideration the different information content in the different training data, small synthetic training patches cover only local structures whereas large patches or 3D slices cover a large region including patient anatomy as well as background information. A model trained with small image patches therefore focuses on local information only and mostly neglects large gradients between background and patient anatomy. However, this large gradients could be important for the transfer on patient data. Since the synthetic data are based on a digital phantom, they display less organ structures in form of e.g. vessels, than real data sets. Thus, real and synthetic patient data mostly differ in terms of local structure, but display both similar structures on a global scale (such as the

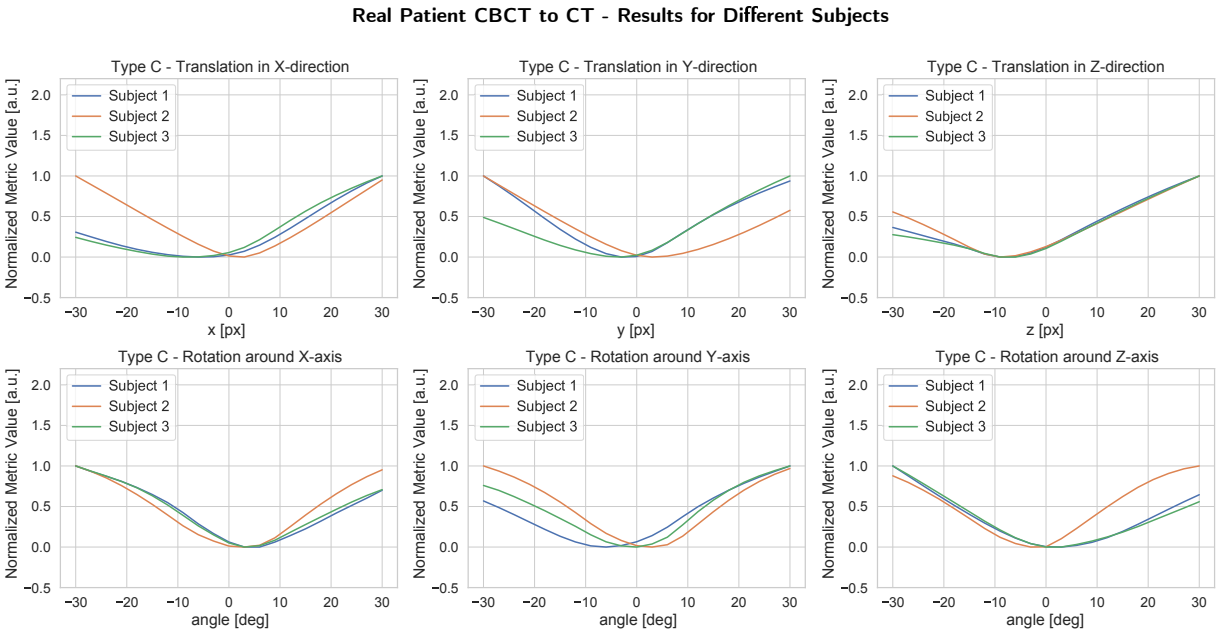


Figure 5.25: Parametric space obtained for the evaluation of different patient CBCT and CT data pairs using the SDM trained with a patch size of $150 \times 100 \times 150$ pixel for translation and rotation transformations.

margin between patient and background). Therefore, models trained with patch sizes larger than $100 \times 100 \times 100$ pixel are preferable for the transfer on real patient data. Since the SDM model trained with *Patches 130* yields a smooth parametric cone with clear minimum, this model is used for all following studies.

Next, the applicability of the SDM is investigated for CBCT/CT data pairs of three different patients, denoted as 'subjects', as shown in Figure 5.25. For all three subjects, the SDM yields a smooth parabolic parametric landscape for translation and rotation transformations with a significant minimum which can be used for the optimization during image registration. Thus, these findings indicate a general transferability of the metric for various real data sets.

At last, the performance of the SDM metric on patient data is compared to traditional multimodal similarity metrics for translation and rotation transformations. The results are shown in Figure 5.26 and Figure 5.27, respectively. For all transformation types, the SDM metric yields a similar position of optimal alignment as the traditional AMMI metric with and without mask. Moreover, the parametric shape obtained for the SDM metric is in good correspondence to the shape obtained with the AMMI metric. This shows, that the metric enables a similar registration accuracy than traditional methods while providing a faster computation time of the metric value, once the SDM model is trained. The SDM metric even outperforms the NGF metric, which yields parametric landscapes with a high amount of local extrema and shifted minima with respect to the optimal alignment of the ground truth data sets situated at position 0.

In summary, the result show that a siamese network is able to learn similarity for various modality combinations as well as morphologies. In addition, the findings presented in this section indicate that it is possible to train the metric using synthetic image data and transfer the model to estimate similarity on real patient data.

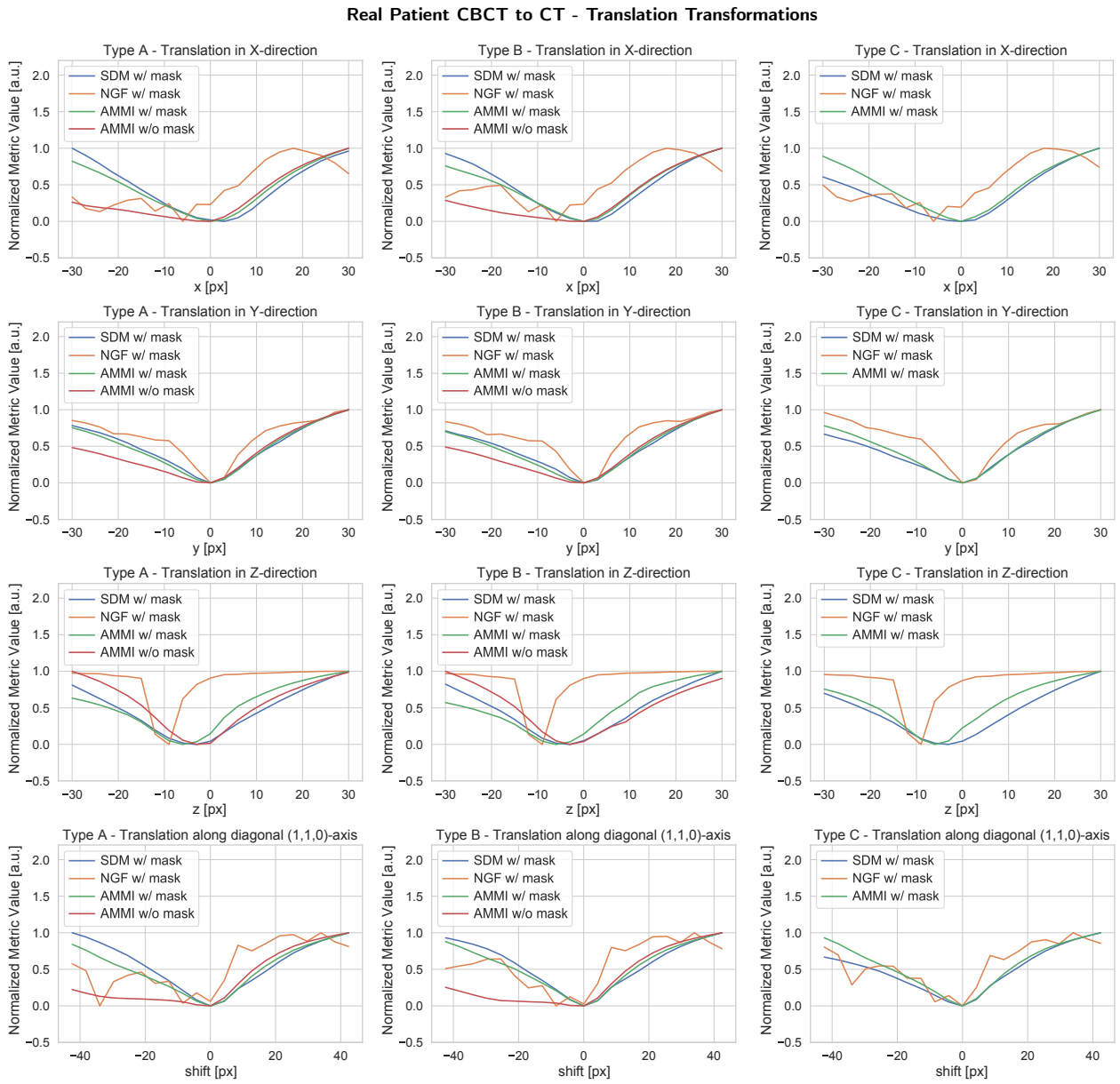


Figure 5.26: Parametric space obtained for a pair of patient CBCT and CT data using the learned SDM, a NGF metric and a traditional AMMI metric with and without the use of an image mask for translation in x-,y-,z- and diagonal (1,1,0)-direction. The metric values were projected in a range from 0 - 1 to enable a comparison between the different metrics.

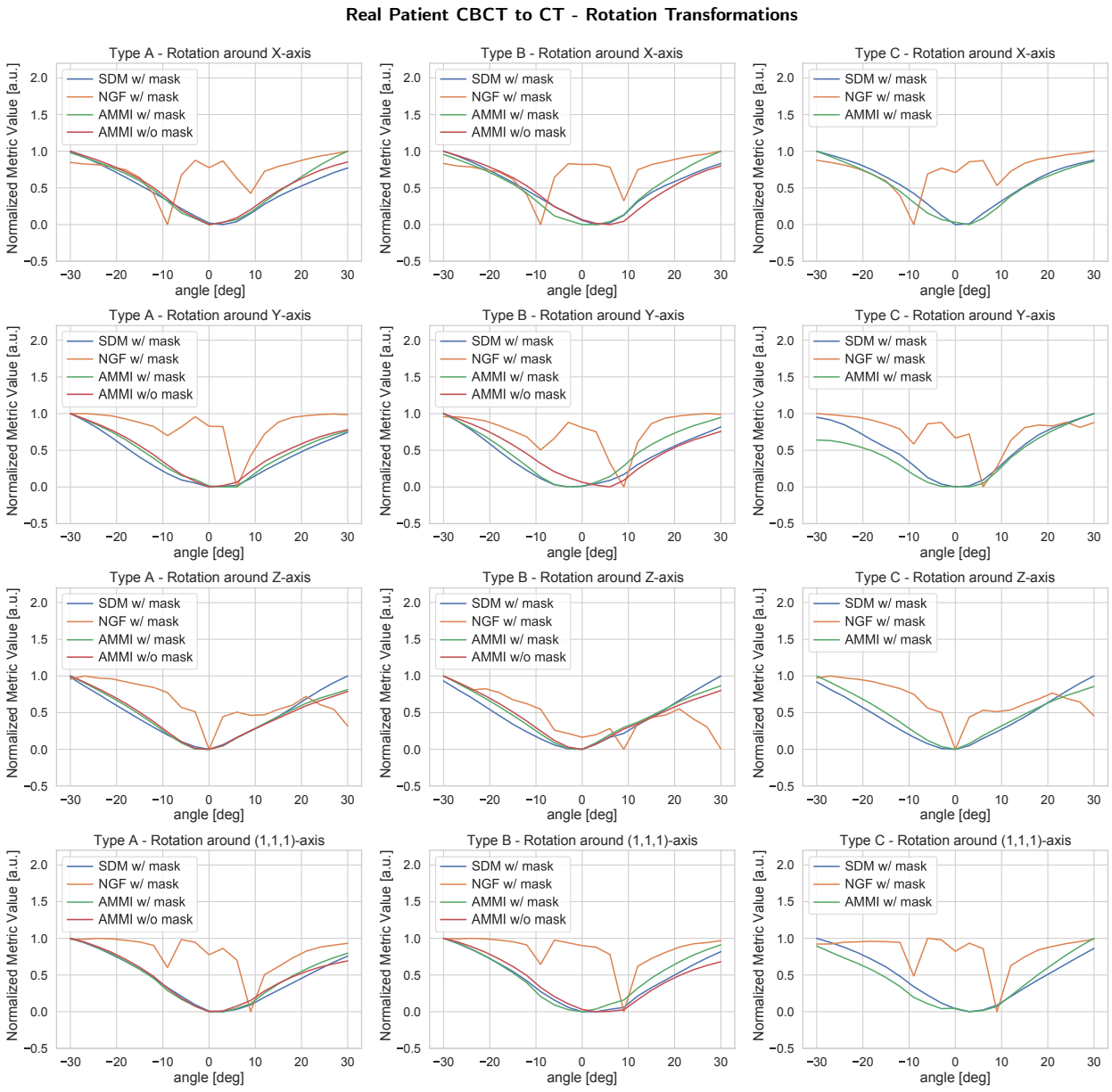


Figure 5.27: Parametric space obtained for a pair of patient CBCT and CT data using the learned SDM, a NGF metric and a traditional AMMI metric with and without the use of an image mask for rotation around the x-,y-,z- and diagonal (1,1,1)-axis.

5.4 End-to-End Image Registration Learning

The last part of the thesis focused on the training of a neural network to learn an entire end-to-end registration process. In total, 24 models were trained using different mono-and multimodal input data sets as wells as loss functions. The registration performance of all models is evaluated by sampling the capture range of the method using the evaluation methodology for rigid registrations presented in Section 4.2.2. The evaluation is divided into three parts, whereas the first part investigates the performance of trained registration models for monomodal registration tasks and the second part focuses on models for multimodal image registration. In the last, the transferability of the models trained with synthetic image data for the application on real patient data is investigated.

5.4.1 Monomodal Registration Learning

The monomodal registration models were trained using two different data sets, corresponding to T1-MRI brain data and synthetic CT abdominal data which also represent the data sets used for the evaluation of the models.

Affine vs. Deformable Registration Models

The original VoxelMorph network is restricted to deformable registration learning. Therefore, the first alteration of the original network was represented by an extension to learn an affine image registration process. Since deformable registration methods are intrinsically restricted to capture local image deformations, it is expected that the original network is not suitable to correct large global image displacements in contrast to the extended affine model.

The results obtained for both registration types using all implemented monomodal loss functions are shown in Figure 5.29 for T1-MRI brain data and in Figure 5.30 for abdominal CT data. For both data sets, the results confirm the assumption that deformable registration models are unable to correct global image displacements, yielding a mean TRE (mTRE) that is only slightly smaller than the mTRE before image registration (see 5.29b and 5.30b). The only exception is represented by the deformable registration model based on a supervised DFL which yields an improved image

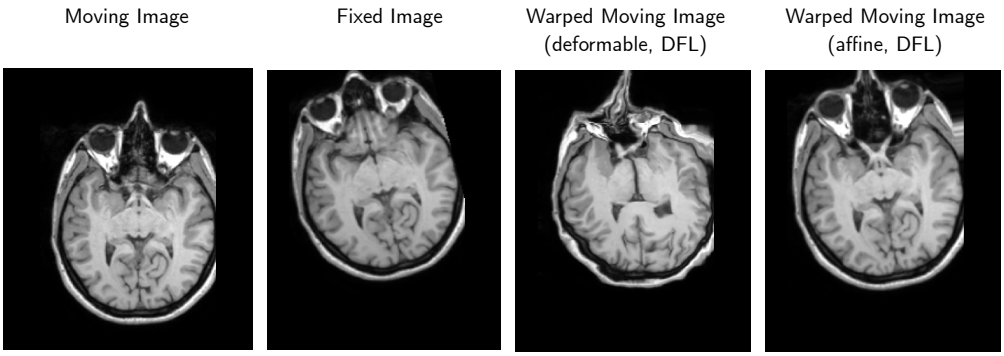


Figure 5.28: Registration example using the affine and deformable registration models trained with the supervised DFL as loss function for the monomodal registration of T1-MRI data.

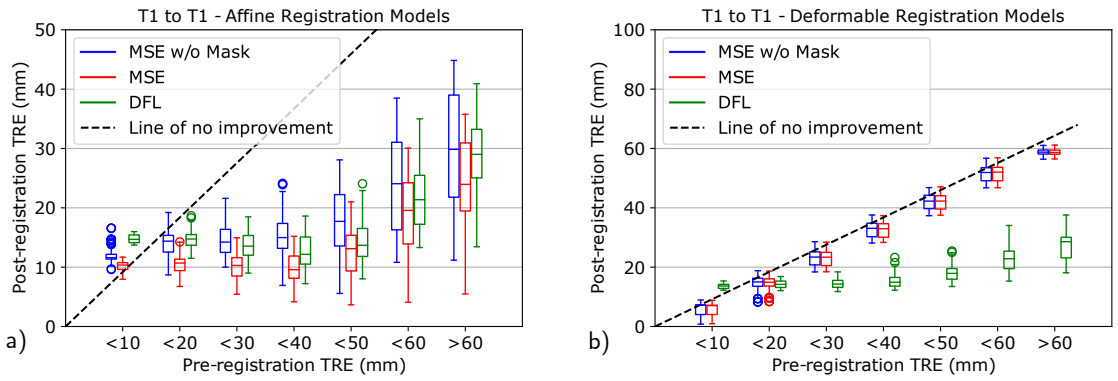


Figure 5.29: **Brain data (T1/T1)**: Monomodal affine and deformable registration of T1 brain MRI using models trained with different loss functions.

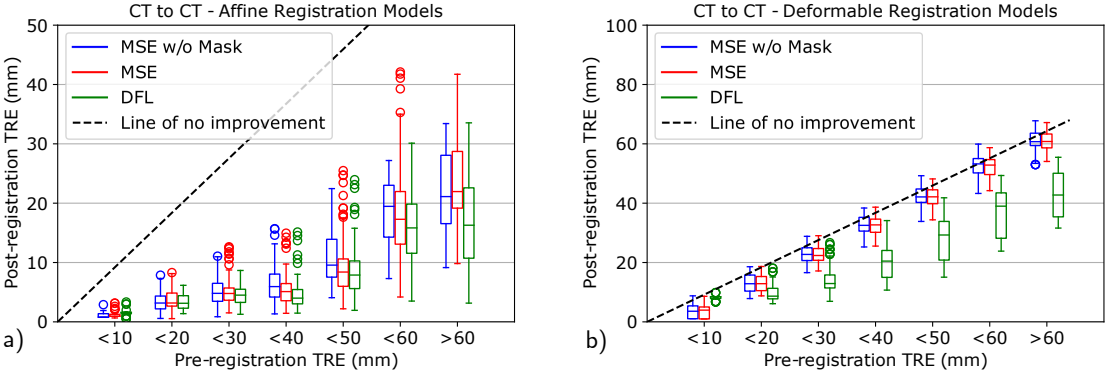


Figure 5.30: **Synthetic data (CT/CT)**: Monomodal affine and deformable registration of synthetic abdominal CT data using models trained with different loss functions.

alignment. Although the deformable registration model succeeds to shift the image in a reasonable direction, its application results in unrealistic tissue deformations due to its flexibility and high number of degrees of freedom. This is exemplary shown for the monomodal registration of T1-MRI in Figure 5.28. In contrast, the affine registration model based on DFL preserves a correct anatomy. Moreover, the comparison of the registration accuracy achieved for the deformable and affine registration models trained with a DFL shows that the latter results in an even lower post-registration TRE.

In general, the affine registration learning models all improve image alignment resulting in a lower post-registration TRE than initial TRE. For image displacements up to 50 mm, the affine registration models yield a mTRE of $10 - 20$ mm for the registration of T1-MRI data (see Fig. 5.29a) and a TRE below 10 mm (see Fig. 5.30a) for the registration of synthetic CT data, depending on the employed loss function. For displacements larger than 50 mm, the post-registration TRE slowly increases for both data sets including a larger standard deviation. The results show that the models are able to correct large image displacements for both data sets. However, the performance for small initial displacements differs depending on the data. For the synthetic CT data, the registration model leads to very small mTREs of ~ 1 mm for initial TREs smaller than 10 mm. In contrast, the models do not improve image alignment for the brain data pairs in this displacement range. This may be linked to the higher amount of available brain data pairs that were used for training

the model. Since they stem from different patients, the data pairs display variances in morphology. This increases the difficulty to learn a registration potentially leading to a lower accuracy for the fine registration of small displacements.

In addition, it has to be noted that the models were only trained for the correction of translation transformations up to 30 mm. Nevertheless, the results show that the affine models were able to correct image displacements even larger than 60 mm for both data sets. This implies a general transferability to novel problems in form of larger image displacements.

Different Loss Functions

Figure 5.29 and Figure 5.30 also display the results obtained for the registration models trained with different loss functions. In total, three loss functions were used for the training of the monomodal registration models: an MSE loss without the consideration of an image mask during metric evaluation as proposed in the original VoxelMorph network, an MSE loss with image mask as well as a supervised deformation field loss.

In contrast to the MSE-based losses which rely on the similarity estimation between the fixed image and the warped moving image, the supervised DFL forces the network to learn the generation of a deformation field that is similar to the known ground truth deformation field. Therefore, the model based on a supervised DFL represents the only model that is able to improve the image alignment for a deformable registration as shown in Fig. 5.29b and Fig. 5.30b. However, since the application of deformable registration models for the registration of affinely transformed image pairs leads to unrealistic image deformations, the comparison of different loss functions focuses on affine image registration applications.

The results obtained for the affine registration models trained with T1-MRI data (see Fig. 5.29a) and synthetic CT data (see Fig. 5.30a) show that all three loss functions represent appropriate measures for monomodal registration learning. All affine registration models yield a mTRE below 20 mm for the T1 brain data and below 10 mm for the synthetic CT data for initial image displacements up to 50 mm. Especially the results obtained for the brain data set show the importance of the use of an image mask during registration. Although all models trained with different loss functions yield similar registration accuracies for small image displacements, the MSE loss without an image mask results in a significantly larger post-registration TRE for large image displacements than the MSE with mask. The consideration of an image mask during the estimation of the loss ensures that only image regions containing valid image information contribute to the loss value. Thus, for the T1-MRI brain data, the lowest mTRE after registration is achieved by using an MSE loss with consideration of a binary image mask. The experiments for this data set have been performed for multiple data pairs, but due to reasons of clarity of presentation, only the results for one data pair are shown. The remaining results are in good agreement with the presented findings and are shown in Appendix 7. For the synthetic CT data, both the MSE loss with and without mask result in similar registration accuracies, whereas the MSE with mask yields a slightly higher registration accuracy. The best loss function for the monomodal CT registration is given by the DFL. However, the requirement of additional information in form of a known deformation field does

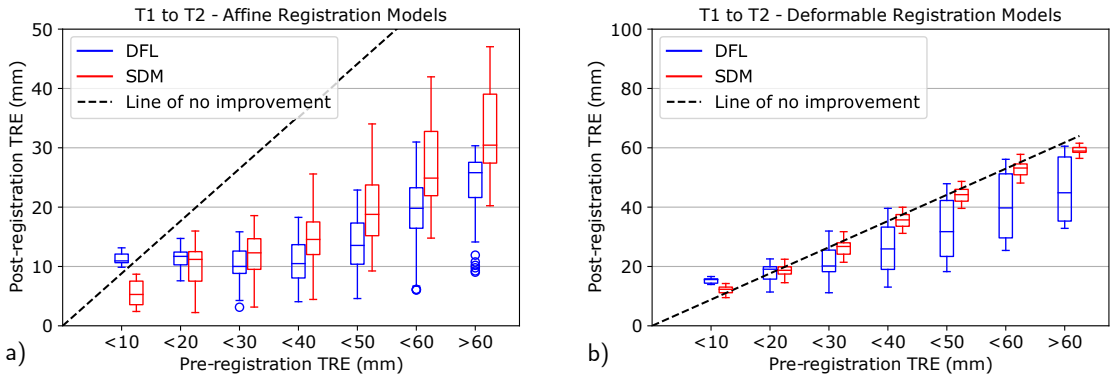


Figure 5.31: **Brain data (T1/T2)**: Multimodal affine and deformable registration of T1 and T2 brain MRI using models trained with different loss functions.

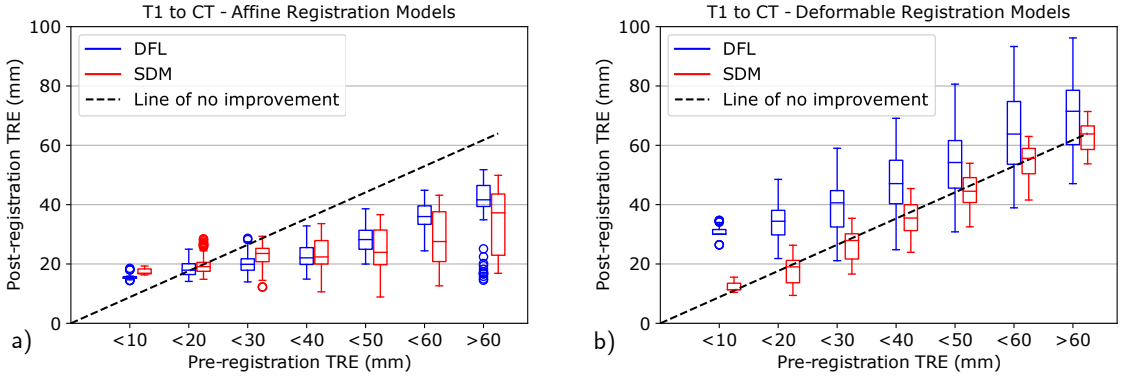


Figure 5.32: **Synthetic data (T1/CT)**: Multimodal affine and deformable registration of synthetic T1 to CT data using models trained with different loss functions.

not advocate the use of the DFL loss, since the unsupervised MSE loss leads to similar registration accuracies, especially for small image displacements below 50 mm.

5.4.2 Multimodal Registration Learning

In addition to the integration of novel monomodal loss functions, two multimodal loss functions were implemented to enable a training of the network for the registration of multimodal image data. The first loss corresponds to the supervised DFL loss. Since this loss only relies on a comparison of deformation fields and not on intensity values, it is suitable for mono- as well as multimodal registration applications. As alternative, the SDM which was discussed in detail in Section 5.3.2 is incorporated as multimodal loss function. With these two losses, affine and deformable registration models were trained for the registration of T1 to T2 brain MRI, synthetic abdominal T1-MRI to CT data as well as synthetic abdominal CBCT to CT data.

Affine vs. Deformable Registration Models

The results obtained for the evaluation of these multimodal models are shown in Figure 5.31 for the registration of T1/T2-MRI brain data, in Figure 5.32 for the registration of T1/CT and in Figure

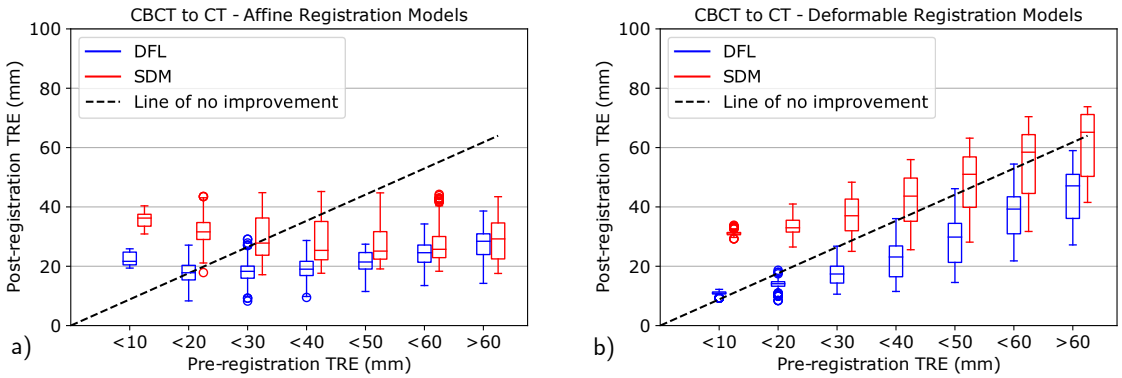


Figure 5.33: **Synthetic data (CBCT/CT)**: Multimodal affine and deformable registration of synthetic CBCT to CT data using models trained with different loss functions.

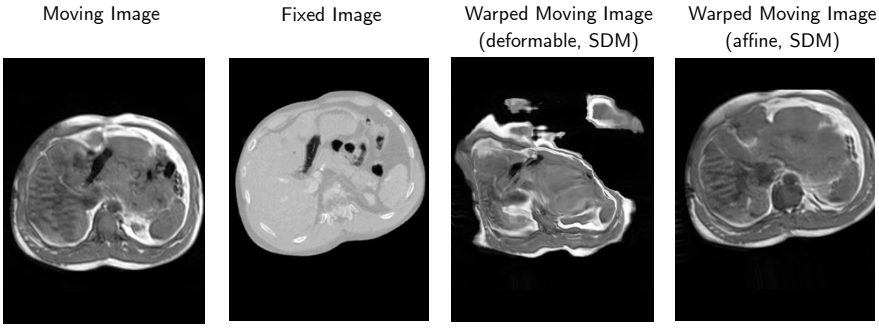


Figure 5.34: Registration example using the affine and deformable registration model trained with the SDM as loss function for the multimodal registration of synthetic T1 to CT data of the abdomen.

5.32 for the registration of CBCT/CT. As for monomodal registration tasks, the results show that the deformable VoxelMorph models are unable to successfully register affinely transformed image data. The results obtained with the deformable registration model trained for the registration of T1- to T2-MRI brain data shown in Figure 5.31 are in good correspondence with the results obtained for the use of monomodal deformable registration models, since their application results in a post-registration TRE similar to the initial TRE. However, the results obtained for the registration of synthetic abdominal data show that, depending on the employed similarity measure, the post-registration TRE may even be larger than the initial TRE for the use of deformable registration models. A possible explanation is given by the different amount of structural information contained in brain and abdominal image data. The T1- and T2-MRI scans are characterized by a lot of texture due to the morphology of the brain. In contrast, the structural information in abdominal data is rather limited since the abdominal organs display a high degree of homogeneity. This is further enforced by the fact that synthetic scans were used as training data which are generated using the segmentation-based digital XCAT phantom which inherently contains less structural information than real patient data. The deformable registration of homogenous image regions represents an ill-posed problem and is generally a challenging task, possibly leading to the increased post-registration TRE observed for the synthetic data pairs.

The performance differences observed for the models trained with different data sets also manifests itself for the evaluation of affine registration models. The registration accuracy obtained for the

affine registration of T1- to T2-MRI is shown in Fig. 5.31a, and for T1 to CT and CBCT to CT in Fig. 5.32a and Fig. 5.33a, respectively. The results show that the magnitude of the post-registration TRE highly deviates for models trained with different data sets. Concerning the affine registration of T1- to T2-MRI data, the registration models yield a mTRE smaller than 20 mm for image displacements up to 50 mm which slowly increases for larger displacements. The mTRE obtained for the registration of T1 to CT and CBCT to CT is significantly larger ranging between $18 - 40$ mm depending on the employed loss function and modality combination for the same range of initial image displacements (note the different y-axis used for brain MRI and synthetic abdomindal data). As for deformable registration, this may be caused by differences in structural image information.

However, the results obtained for all data sets show that the affine registration models generally achieve an improvement of the image alignment for large image displacements. As for monomodal registration tasks, the extension of the VoxelMorph network to affine registration results in a higher registration accuracy as deformable registration models while preserving the morphology as shown in Figure 5.34.

Different Loss Functions

As for monomodal registration models, the comparison of models trained with different loss functions is limited to the application of affine registration tasks, since the use of deformable registration models generally leads to unrealistic image deformations.

Concerning the range of large image displacements (larger than 40 mm), the results show that models trained with both loss functions, the DFL and the SDM loss, lead to an improvement of image alignment. For all data pairs, the standard deviation of the registration accuracy obtained with the DFL-based model is significantly lower than the standard deviation obtained using the SDM-based model. This indicates a higher robustness of the models trained with the DFL loss. In terms of registration accuracy, the DFL-based model outperforms the SDM loss for the registration of T1/T2-MRI as well as CBCT/CT data. In contrast, the SDM loss leads to a lower post-registration TRE for the registration of T1-MRI to CT.

The performance of the different models for small pre-registration TREs highly depends on the image data used for training and evaluating the network. Whereas the model trained with the SDM loss improves the image alignment for small image displacements below 10 mm for the registration of T1/T2-MRI, it leads to an even higher post-registration TRE for the registration of CBCT to CT data for displacements below 30 mm. This indicates the requirement of further optimization of the SDM, so that it leads to stable registration results for all modality combinations. In contrast, the performance of the DFL is similar for all data pairs, indicating a general applicability of the method.

In general, the registration accuracy of the multimodal models is limited for small image displacements below 10 mm. This may be related to the quality of the ground truth image data used for training the models. Since the brain data corresponds to real patient data, there might still exist small displacements despite of the registration performed during preprocessing of the data. As for

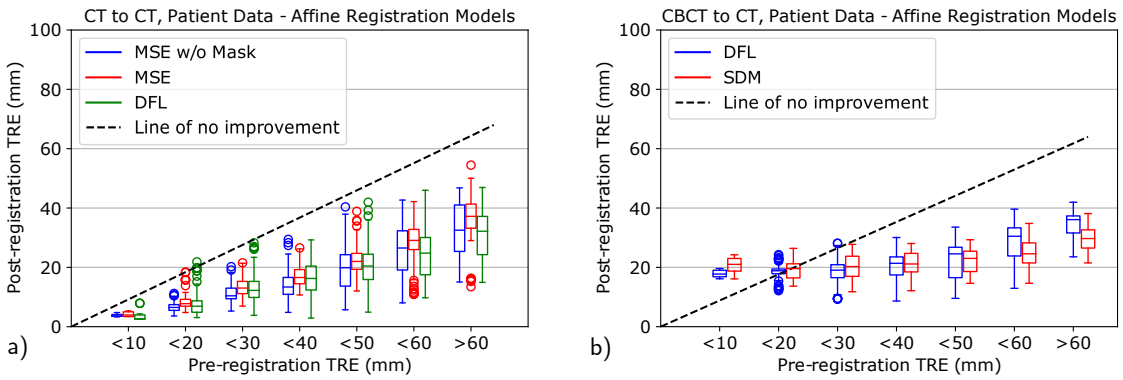


Figure 5.35: **Real patient data (CT/CT, CBCT/CT)**: Monomodal and multimodal affine registration of real patient data using models trained with synthetic CBCT to CT data based on different loss functions.

the generation of the synthetic data, the application of the CycleGAN might result in blurring on the organ edges (as discussed in Sec. 4.2.1). These small morphological deviations in the ground truth data pairs translate to the accuracy that is achievable by the registration models.

5.4.3 Transfer to Patient Data

At last, the performance of the registration models trained with synthetic image data is investigated for the application on exemplary data pairs of real patient data. This transfer included the evaluation of the affine registration models trained for the monomodal registration of abdominal CT data as well as the multimodal registration of CBCT to CT data. To establish a comparability of the synthetic and real patient data, the intensity values of real CBCT and CT data were rescaled using *SimpleITK*, so that the intensity distributions are similar to the intensity distributions of the data pairs that were used for training the models. The evaluation was performed for three data pairs. Here, the results for one pair are exemplary shown whereas the remaining results are in good correspondence with the findings presented here and are shown in Appendix 7.

Figure 5.35a displays the registration accuracy obtained for the monomodal registration models trained with different loss functions. The results show that the application of the monomodal affine registration model on real patient data leads to an improvement of the image alignment, since all models yield a post-registration mTRE smaller than the initial mTRE. However, the obtained registration accuracies are characterized by large standard deviations. As for the evaluation on synthetic image data, the models trained with different loss functions yield very similar registration results, whereas the model trained with DFL yields the highest registration accuracy for large image displacements.

The results obtained for the application of the multimodal registration models trained with synthetic CBCT/CT data on real CBCT/CT data are shown in Figure 5.35b. Both models trained with different loss functions lead to a clear improvement of the image alignment for initial image displacements larger that 20 mm, whereas the model which was trained using the unsupervised SDM slightly outperforms the DFL-based model in terms of registration accuracy. Comparing these results to the results obtained for the application on synthetic image data, it can be seen that

the model performs very similar on both data sets.

In summary, the findings presented in this section suggest the transferability of the models trained for the registration of synthetic image data for applications based on real patient data. However, the transfer is linked to some requirements, such as a preprocessing of the patient data so that the underlying intensity distributions approximately correspond to the intensity distribution of the image pairs used for training the network. Moreover, the VoxelMorph network can only handle input images having the same size, therefore the preprocessing should not only include the rescaling of intensity values but also a resampling of the image dimensions if necessary. The results show that the models successfully achieved an improvement of the image alignment for monomodal as well as multimodal registration.

6

Discussion

In this chapter, the findings presented in the previous chapter will be summarized and their implications for the field of medical image registration are discussed. The chapter is divided into three parts. The first part includes a review of the developed evaluation methodology for linear and non-linear multimodal registration approaches and its potential applications. The second part provides a discussion of the performance of the novel image similarity measures proposed in this thesis. The last part of this chapter focuses on the discussion of the performance of developed models for mono- and multimodal image registration learning using a Deep neural network.

6.1 Evaluation of Linear and Nonlinear Image Registration

The developed evaluation methodology has been used for the optimization and performance comparison of methods for the multimodal registration of T1-MRI to CT data as well as CBCT to CT data. The different registration methods are characterized by variations such as different similarity measures, fixed image masks, the number of resolution layers or specific features such as the physical grid spacing of a deformable spline transform.

Concerning the registration performance based on different similarity measures, the results for linear and non-linear registration methods are consistent. For both registration types, the results indicate that similarity measures based on mutual information yield a high accuracy for the registration of CBCT/CT and T1-MRI/CT abdominal scans. Within the frame of this thesis, two different metrics based on mutual information were investigated, namely Advanced Mattes Mutual Information (AMMI) and Normalized Mutual Information (NMI). The results obtained for the comparison of these metrics suggest that NMI leads to a minimally lower registration error than AMMI and is therefore favorable for the registration of multimodal abdominal image data.

Another correspondence for both registration types occurred for the estimation of the optimal number of resolution layers in a multistage registration. The experiments show that more than three resolution layers do not improve registration accuracy, but only increase computation time.

A significant difference for both registration types is the behavior observed for the application of fixed image masks during registration. Deformable registration methods improve by increasing the image region in which the similarity measure is estimated during registration. Thus, the

highest registration accuracy is achieved without an image mask. This can be explained by the fact that deformable registration methods are very flexible. The application of a fixed image mask restricts the image region considered during metric evaluation, which potentially leads to a limitation of local information that is necessary to estimate an appropriate spatial deformation field. This is especially significant for the registration of very homogenous structures, such as the liver. In contrast, linear registration methods benefit from the introduction of a mask. During the experiments for linear registration methods using actual patient data, it became apparent that a fixed image mask is mandatory for the registration of CBCT as moving image to CT as fixed image, to prevent the algorithm from failure. This is mainly due to the fact that rigid registrations rely on global information and are generally intended to correct geometrical displacements of a larger magnitude than nonlinear registration methods. Taking into consideration the intended application scenario of interventional procedures, masks are usually not available in a routine setting. The results indicate that even the introduction of a simple cuboid mask, which could be provided manually with acceptable effort before the intervention, can improve the registration outcome. Moreover, with the upcoming of deep learning based segmentation approaches, fully automated, robust segmentation of target structures, such as the liver, is available with accuracies that should be sufficient to be used as image masks for registration.

Especially the comparison of registration methods with and without the consideration of image masks revealed that the registration of CBCT to CT data is a very difficult task. This is mainly attributed to the often low image quality of CBCT data that is characterized by a high level of noise as well as a limited FOV due the image acquisition and reconstruction. The difference in the FOV represents a challenge, since CBCT data often only covers the anatomical target region, such as the liver, whereas a CT generally covers the entire abdomen. This leads to a significant difference in size of both images which can additionally complicate the registration procedure. A possible improvement is given by the performance of an initialization procedure which roughly prealigns the image volumes before the actual registration. The experiments performed for liver scans in this thesis show that the best initialization for CBCT to CT data is given as superposition of the geometrical centers of the target structure, such as the liver in the context of this thesis. In addition, deviations in terms of grey value distributions in CT and CBCT data contribute to the registration problem since there exists no defined relation between the intensity values in both modalities. Although results presented in this thesis show that monomodal similarity measures relying on cross-correlation may be appropriate for the registration of CBCT to CT, the results obtained with a mutual information based metric outperform monomodal metrics. Thus, the results presented in this thesis suggest the application of a multimodal rather than a monomodal similarity measure for an accuracte registration of CBCT to CT data.

In general, the developed evaluation methodology for rigid and non-linear registration methods enables a fast comparison and optimization of different techniques for the multimodal registration of abdominal scans. In contrast to existing evaluation methods, the presented method does not rely on overlap-based accuracy measures which represent arguable quality criteria for registration evaluation [112]. Instead, the developed evaluation method uses a point-based accuracy estimation which employs a dense grid of landmarks to estimate registration accuracy within the liver.

Due to the modular setup of the *SimpleElastix* toolkit, an easy implementation of a multitude of different registration methods as well as the adjustment of various registration parameters is possible. In addition, the component performing the actual registration in the evaluation framework can be replaced by any registration method, thus also enabling the evaluation of novel registration approaches. This is demonstrated by the use of this novel evaluation methodology to benchmark the performance of an algorithm that relies on a neural network to learn an entire end-to-end registration developed within the frame of this thesis (see Section 4.4). Another advantage of the methodology developed for the evaluation of rigid registration methods is the fact, that it can be easily extended for the estimation of registration accuracy for alternative morphologies, such as e.g. brain data. To realize such an extension, ground truth registered data is required as well as segmentation masks for the target structures in which the registration accuracy should be estimated. By using this data as novel input to the methodology, it is possible to apply the registration accuracy for alternative morphologies or data sets. However, such an extension to alternative morphologies is not as simple for the evaluation developed for non-linear registration methods, due to the lack of non-linearly deformed ground truth data. An extension could potentially be realized by applying artificial deformable transformations on prealigned ground truth data, and using these generated deformation fields as input to the methodology. However, artificial non-linear transformations rarely depict realistic deformations of organs and soft tissue.

Nevertheless, due to its wide range of application possibilities and easy modular setup, the developed evaluation methodology represents an important contribution to benchmarking various methods for an optimized registration of abdominal image data.

6.2 Novel Similarity Metrics

6.2.1 Similarity Metric based on *Histograms of Oriented Gradients*

The evaluation of the novel similarity metric based on the comparison of 3D-HOG descriptors was divided into two parts. The first part focused on the sampling of the parametric space of the metric for different translation and rotation transformations using pre-interventional CT data and intra-interventional CBCT data of three different patients. For the second part, the novel metric was successfully implemented in an entire registration algorithm for CBCT/CT which was evaluated using the methodology for rigid registration methods discussed in the previous section.

The evaluation of the parametric spaces showed, that the HOG-based metric yields a clear global minimum for translation and rotation transformations which can be identified during an optimization process. However, the metric displays tendencies to local minima in the parametric space for large initial displacements eventually causing misregistrations depending on the utilized registration optimizer.

Nevertheless, the results obtained for registration evaluation for translation displacements up to 30 mm and rotation displacements up to 30° indicate that the HOG-based registration algorithm is able to successfully correct medium image displacements. These reach up to 14 mm translations and 10° rotations. The results show that the probability for successful registrations in this range

is comparable to the probability obtained with an NCC-based algorithm. The comparison of the HOG-based metric to the NCC showed that the HOG metric is generally able to compete with NCC in terms of the registration accuracy. For the data used in this evaluation, the novel metric performs slightly inferior for the registration of rotational image displacements, but displays a higher performance for the registration of translation transformations. Further studies have to show if this is a general behavior of both registration metrics, or if this circumstance is only related to the used image data.

In general, the patch-wise computation of the HOG-based metric provides a high level of robustness to noisy image data and illumination changes, since it is not affected by the inevitable variety of corresponding pixel-to-pixel intensities. Another advantage of the HOG-based similarity metric is the fact that it has the potential to be usable not only for monomodal but also for multimodal registration applications, since the metric exclusively relies on gradient information. Distinct gradients in medical images are often related to organ margins or sudden structural changes which are often visible in images of different modality. Moreover, the feasibility study done in this work showed that the new metric can compete with established similarity metrics in terms of registration accuracy for the challenging task of registration of CBCT to CT data.

However, a major drawback of the HOG-based similarity measure is the fact that the extraction of a 3D-HOG descriptor may be very time consuming depending on the size of the input images. In cooperation with project partners from the Department of Computer Assisted Surgery, Medical Faculty Mannheim, University of Heidelberg, a parallelized implementation of the extraction algorithm has been implemented, reducing the computation time from 250 seconds to 1.1 seconds for an image volume of $512 \times 512 \times 188$ pixel [195]. But the complexity of the HOG estimation still makes the HOG-based similarity metric unsuitable to be used for clinical applications requiring fast computation. To make the algorithm usable in a clinical context, further adjustments to improve the performance of the metric for rotation transformations would be required. Moreover, the influence of different metric parameters, such as chosen histogram bin size, cell size or block size, on the shape of the parametric space should be investigated further.

6.2.2 Similarity Metric based on a Siamese Neural Network

The performance of the learned similarity metric based on a siamese neural network was characterized for different modality combinations and morphologies by sampling the parametric space for translation and rotation transformations.

Concerning the application of the learned similarity metric for different morphologies, the results obtained for brain and abdominal data show that the metric is able to define an optimum in the parametric space for all investigated patient morphologies. The position of the optimum was in good agreement with the position identified by traditional similarity measures such as AMMI and NGF. This indicates that the neural network is able to distinguish between 'similar' and 'dissimilar' images of different anatomical data. However, the results indicate that the performance of the metric is affected by the complexity of the morphological structures displayed in the images. Although the evaluation for both morphology type yields smooth parametric space landscapes with few local

extrema, the parametric space obtained for complicated structures such as the brain displays a decreased smoothness compared to the result obtained for abdominal image data.

The performance of the Siamese Deep Metric was additionally evaluated for different modality combinations, including intra-modal T1/T2-MRI as well as multimodal CBCT/CT and T1 MRI/CT data. The learned similarity results in smooth parametric cones for all modality combinations which indicate a general usability for image registration purposes. This behavior could be confirmed for the evaluation of several data pairs per modality combination.

Instead of a general usability of the metric for universal image data, the Siamese Deep metric has to be trained using image data specified for an intended task. This is simultaneously an advantage and a drawback of the method. It can be advantageous since the metric is trained and optimized to estimate the similarity between very specific data. This is exemplary demonstrated for the comparison of the novel Siamese metric to the traditional NGF metric. This comparison showed that the Siamese metric outperforms the NGF metric for certain data types, such as e.g. the MRI brain data pairs, by providing a smoother parametric space landscape which is beneficial for image registration. However, the training of the siamese network to learn such a similarity estimation can be time consuming and requires registered multimodal ground truth data. This is a major drawback of the method, since this type of data has to be tediously obtained first or may not exist at all due to the acquisition process of different image modalities.

For one type of experiments performed in this thesis, this limitation is overcome by training the network on synthetic image data. Since the aim of the work presented in this thesis is the development of novel methods for the registration of real pre-and intrainterventional patient data, the transferability of a learned Siamese metric model trained with synthetic image is investigated by applying the model on real patient CBCT/CT data. The results show, that this transfer is generally feasible and the parametric space landscapes obtained for the Siamese Deep metric are in good agreement with the results obtained for traditional multimodal similarity measures such as AMMI. This general transferability of the metric can be very beneficial, since the main advantage of such a learned similarity measure is the fast computation time of metric values (in the range of subseconds depending on the image size), once the model is trained. The implementation of such a metric in an entire registration algorithm could therefore significantly increase the computation speed of the registration, since the metric estimation traditionally corresponds to the most time consuming component of the iterative procedure. Another benefit is the fact that it is possible to integrate the trained Siamese metric as multimodal loss function in another neural network as will be discussed in the next section. This is important, due to the promising results obtained for novel approaches relying on DL in the field of medical image processing. However, up to now, only very few loss functions applicable for the similarity estimation of multimodal image data exist for the use in DL frameworks such as *Keras* or *TensorFlow*.

The results generally show, that it is possible to learn a similarity measure and successfully transfer the trained models for the similarity estimation on data sets with comparable intensity distributions. The fast computation time as well as the smooth parametric landscapes make these learned Siamese similarity measures an interesting alternative to traditional similarity measures, especially for the development of novel approaches requiring a multimodal similarity metric in the context of DL.

Comparing the findings obtained for the novel Siamese metric to the findings concerning the HOG-based similarity metric, it can be stated that the Siamese metric displays a higher robustness for large rotations. Moreover, the computation time of the metric value is several magnitudes faster than for the HOG-metric. However, a significant advantage of the HOG-metric is the fact that it does not rely on a training requiring ground truth registered image data. In general, the HOG-metric is applicable to various data sets with the sole requirement of distinct image gradients. Both metrics represent appropriate alternatives for the similarity estimation of multimodal data, which up to now is dominated by mutual information based metrics.

General Remarks

Aside from the performance evaluation of the Siamese Deep metric, the results obtained by sampling of the parametric space for different data sets using traditional similarity metrics provided additional important findings. For all experiments, three different metric evaluation cases were distinguished which differ by the position of optimal image alignment and the application of image masks. For one of these cases, the optimal position of image alignment corresponds to a position with a slight offset and not to the position of maximal overlap of the images. The results obtained for the sampling of the parametric space using AMMI with and without mask for this specific case showed that the introduction of a binary image mask can be essential to identify the correct position of image alignment. In general, the introduction of a binary image mask during similarity estimation restricts the regions in both images taken into account for the computation of the metric values to regions containing only valid image information. Non-valid regions containing default pixel values which are created during image resampling processes after geometric transformations are discarded when using an image mask. Hence, only image regions containing actual image information are compared for the similarity estimation. Without a mask, the default region of one image may be compared to a valid region of another image, leading to incorrect metric values, as shown for the results of the AMMI metric with and without mask. Up to now, the use of image masks is not always a part of image registration methods especially for approaches based on neural networks. However, the results obtained within the frame of this thesis indicate a high importance of these masks to improve registration accuracy.

6.3 End-to-End Image Registration Learning

Deformable registration methods are generally limited to correct local image displacements, such as e.g. organ deformations due to respiratory motions. However, for many applications in medical image processing, it is necessary to correct larger, global image misalignments. Therefore, the first modification of the VoxelMorph network is represented by learning the entries of an affine geometric transform matrix. This matrix builds the basis for the calculation of a 3D deformation field that is used to warp the moving image during registration learning. The results show that this extension successfully enables the learning of an affine registration process. The direct comparison between the deformable and affine registration models confirm the assumption that deformable models are unable to correct global displacements, whereas the affine registration models improved image

alignment for all investigated data sets and modality combinations. As expected, the evaluation showed that affine registration learning generally preserves anatomy whereas the application of deformable registration models lead to unrealistic tissue deformations.

Moreover, the results obtained for monomodal and multimodal registration tasks show that the affine registration models are able to correct image displacements larger than the image displacements used for training the network. Although this circumstance leads to an extended capture range of the method, the training parameters should generally be chosen similar to the parameters of the intended registration task to ensure an optimal registration performance. For both multimodal data sets, the brain MRI data as well as the abdominal image data, the models do not improve image alignment for small displacements below 10 mm. This is most likely related to the accuracy provided by the multimodal ground truth data pairs. Therefore, future investigations should focus on an improvement of the ground truth generation processes.

In a clinical context, an ideal registration algorithm would be able to correct global and local image displacements. Therefore, an additional variation of the network could be realized by implementing a multistage registration which combines a global affine registration to generate an initial image alignment with a deformable registration to correct local image displacements.

In addition to the extension to affine registration learning, novel loss functions were investigated to train the network for mono-and multimodal registrations. Concerning monomodal registration learning, three loss functions were employed including an MSE without the use of an image mask during metric evaluation, an MSE with image mask and a supervised DFL that is based on the comparison between deformation fields. The results show that the affine registration models trained with all three losses lead to an improvement of the image alignment and are suitable to be used as loss functions for training the registration network. However, comparing the registration accuracies obtained using the three loss functions, it can be stated that the MSE loss without consideration of an image mask performs slightly inferior to the other two loss functions for the evaluation of all data sets. This confirms the observations stated in the previous Section 6.2, that the use of an image mask during metric evaluation leads to an improvement of the registration accuracy, since only valid-image regions contribute to the similarity estimation of the images. Both the model trained using the MSE loss with image mask and the model trained using the supervised DFL yield very similar registration accuracies. But since the DFL loss relies on additional information in form of a known deformation field, the unsupervised MSE loss with image mask represents the most suitable loss function of all three options for monomodal registration learning.

Concerning multimodal registration learning, two loss functions were employed: Since the DFL relies on the comparison of the known deformation field applied during training and the deformation field generated by the neural network, it is applicable for mono- as wells as multimodal registration tasks. In addition, the SDM is integrated as an alternative multimodal loss function.

The results show that the models trained with both multimodal loss functions lead to an improvement of image alignment for large displacements for all modality combinations. However, the performance for small image displacements highly varies depending on the employed image data. Whereas the SDM-based models succeed to improve image alignment for small displacements below 20 mm for the registration of T1/T2-MRI, it decreases the registration accuracy for the registration

of CBCT to CT data as well as T1/CT data. In turn, the DFL-based models decrease registration accuracy for small image displacements below 10 mm for all modality combinations, but outperform the SDM-based models in terms of registration accuracy for larger displacements for the registration of T1/T2-MRI and CBCT/CT. For all data pairs, the DFL lead to a lower standard deviation of the registration accuracy compared to the SDM-based models, indicating a higher robustness of the method. Thus, DFL does not necessarily offer an advantage for monomodal registration learning but after further optimization could represent a useful alternative for multimodal registration learning. In general, the result show that both multimodal loss functions are able to improve image alignment for the registration of multimodal image data.

However, further optimization is needed to increase the registration accuracy of multimodal models trained with both loss functions for the registration of small image displacements. The main reason for the decreased registration accuracy in this range may be found in the quality of the alignment of the ground truth data pairs used for training the network. Since the MRI brain data stems from real patients, the data sets is characterized by a high degree of morphological variances which leads to an increased complexity of learning an optimal image alignment. Concerning the synthetic data pairs, the generation process using the CycleGAN may lead to slight blurring effects which may also decrease the alignment accuracy in these data sets on a finer mm-scale. Therefore, future investigations should focus on the quality of the ground truth data to improve registration accuracy of the trained registration models. As for now, the results indicate a general usability of the models to generate an appropriate initial registration of the data pairs. Due to the fast computation time of this registration, a reasonable scenario in a clinical context may include the performance of a multistage registration consisting of a model-based preregistration to roughly align the images followed by a traditional registration to increase the accuracy for remaining smaller image displacements.

As last part of the evaluation, the models trained with synthetic image data were evaluated on real patient data to investigate the transferability of the method. The results show that the models improved image alignment for both, the monomodal registration of CT/CT data as well as the multimodal registration of CBCT/CT data. In general, the transfer is linked to several requirements. These requirements include the preprocessing of the input data so that its intensity distributions resemble the intensity distributions of the training data. Moreover, the network requires input data with an identical image size leading to the requirement of an additional image resampling for image pairs with different size. Nevertheless, the results obtained in this thesis suggest a transferability of the method to data sets with similar image characteristics. To further characterize the transferability, an in-depth study of the tolerance of the method concerning differences in intensity distributions or variations in morphological features may be of interest.

In summary, the results obtained in this thesis confirm the general applicability of a neural network to learn entire monomodal as well as multimodal affine registration processes. Although the results indicate the requirement of optimized training data to improve the performance of the models in terms of registration accuracy, especially for the correction of small image displacements, the networks generally achieved a significant improvement of the image alignment for larger initial image displacements. Thus, end-to-end registration learning may represent an interesting alternative to

traditional methods in the future, since it leads to a significant acceleration of the entire registration process to (sub-)seconds, once the network is trained. As for all multimodal registration approaches relying on Deep Learning, the main challenge of using such a network is its requirement for ground truth training data. The results obtained in this thesis indicate the transferability of the registration models to data sets with very similar intensity distributions as the training data. Therefore, a possible solution to the requirement of ground truth data may be realized by training the network on synthetic data before applying it on real patient data.

7

Summary and Outlook

Image registration methods that are specifically designed for the registration of abdominal image data are only scarcely available. Therefore, the aim of this thesis was the development and optimization of approaches for the registration of multimodal image data of the abdomen. The focus is set on the multimodal registration of abdominal T1-weighted MRI, CT and CBCT data with regard to the use-case of diagnosis and treatment of liver cancer. This work contributes to three different research areas in the field of medical image registration:

1) Image Registration Evaluation and Generation of Ground Truth Data
To evaluate the performance of a registration method, reference data in form of registered image pairs is required which are hardly available, especially for multimodal applications. Therefore, two approaches for the generation of multimodal ground truth data are presented: a pre-registration of actual patient data based on anatomical landmarks as well as the generation of synthetic ground truth data using a neural network. Based on this data, an evaluation methodology has been developed that benchmarks methods based on registration accuracy in the liver. The presented evaluation methodology enables a performance comparison as well as an optimization of rigid, affine and deformable registration methods. It has been used to investigate the performance of different registration methods for the registration of CBCT to CT patient data as well as T1-MRI to CT patient data. The methods are characterized by the variation of multiple parameters including different initialization methods, similarity measures, fixed image masks, the number of resolution layers or specific features such as the physical grid spacing of a deformable spline transform.

As briefly demonstrated in this thesis for brain data, it is possible to extend the evaluation methodology for different morphologies making it a suitable tool to optimize various registration methods for a very specific task. In addition, the integration and optimization of additional registration methods may lead to the generation of a portfolio of methods that are suitable for multimodal abdominal image registration.

2) Novel Similarity Measures
The evaluation results obtained for standard registration methods demonstrated the high impact of the similarity measure on the achievable registration accuracy. Due to differences in intensity distributions, it is especially demanding to establish a similarity estimation between images of different modality. In this thesis, two alternative multimodal similarity measures were proposed.

The first metric relies on a comparison of gradient information in form of *Histograms of Oriented*

Gradients (HOG) in the images. The HOG feature descriptor which was originally developed for 2D image processing applications has been extended for three dimensions and implemented as basis of an image similarity measure. The novel metric was evaluated for the registration of CBCT/CT patient data. The results confirm the assumption that HOG feature descriptors are suitable to be used as basis of an image similarity metric and that the metric can compete with traditional similarity measures in terms of registration accuracy. Due to its patch-wise computation the metric is robust to image noise and illumination changes in the images. By variation of metric parameters, such as the histogram bin size, the cell size or the block size, the metric may be further optimized for specific data types.

As an additional alternative to conventional image similarity metrics, an approach based on multimodal *Deep Metric Learning* using a Siamese neural network was proposed. The evaluation of the second similarity metric focused on the sampling of the parametric space for translation and rotation transformations. Moreover, the applicability of the metric for different morphologies and modality combinations was investigated. The results show, that it is possible to learn a similarity measure for various data combinations. Although the network training requires ground truth image data, the results demonstrate a successful transfer of models trained on synthetic data to real patient data sets. Due to this transferability and its fast computation times, this metric represents an interesting alternative to traditional similarity measures. Future studies should focus on variations of the network architecture as well as an in-depth study of the tolerance of the metric concerning differences between training and test data, such as e.g. differences in intensity distribution.

3) End-to-End Registration Learning

The last part of this thesis focused on the development and evaluation of methods for end-to-end registration learning using a neural network. The network architecture employed for this task is based on the *VoxelMorph* network presented by Balakrishnan *et al.*[6] which is designed for monomodal deformable image registration learning. Within the frame of this thesis, this network has been extended for affine registration tasks as well as for the application on multimodal image data. The extension includes the consideration of an image mask during loss calculation and the integration of novel loss functions, such as the pretrained Siamese metric and a supervised deformation field loss. The performance of the developed registration models was then investigated using different data sets, including the monomodal registration of T1-MRI brain scans and synthetic abdominal CT data pairs as well as the multimodal registration of T1/T2-MRI brain scans, synthetic T1/CT and CBCT/CT scans. The evaluation of the registration models also included the successful transfer of models trained using synthetic data pairs to the registration of real patient data. The results confirm the possibility to learn entire monomodal and, most notably, multimodal registration processes using a neural network with the actual registration being orders of magnitude faster than traditional registration methods. This may be especially relevant for time-critical application scenarios.

In a clinical context, an ideal registration should be able to correct global and local image displacements. Thus, a multi-stage registration which combines the global registration learning developed in this thesis with deformable registration learning provided by the original VoxelMorph network may be beneficial and should be subject to further investigations.

In conclusion, this thesis proposes improved as well as novel approaches for the multimodal registration of abdominal image data. The presented work enables the optimization of existing registration methods in terms of registration accuracy also beyond the use-case of interventional procedures in the liver. Moreover, alternative registration approaches relying on novel similarity measures are discussed as well as the exploitation of Deep Learning for image registration. Especially methods relying on Deep Learning represent impressive alternatives to traditional registration methods, since they bear the potential to yield similar or even higher registration accuracies while performing orders of magnitude faster.

Bibliography

[1] A. H. Foruzan and H. R. Motlagh, "Multimodality liver registration of Open-MR and CT scans," *International Journal of Computer Assisted Radiology and Surgery*, vol. 10, pp. 1253–1267, 2014.

[2] Y. Chen, R. Xu, S.Tang, S. Morikawa, and Y. Kurumi, "Non-rigid MR-CT Image Registration for MR-Guided Liver Cancer Surgery," in *IEEE/ICME International Conference on Complex Medical Engineering*, pp. 1756–1760, May 2007.

[3] Z. Xu, C. P. Lee, M. P. Heinrich, M. Modat, D. Rueckert, S. Ourselin, R. G. Abramson, and B. A. Landman, "Evaluation of six registration methods for the human abdomen on clinically acquired CT," *IEEE Transactions on Biomedical Engineering*, vol. 63, pp. 1563–1572, Aug 2016.

[4] J. B. A. Maintz and M. A. Viergever, "A survey of medical image registration," *Medical image analysis*, vol. 2 1, pp. 1–36, 1998.

[5] J. P. W. Pluim, J. B. A. Maintz, and M. A. Viergever, "Mutual-information-based registration of medical images: a survey," *IEEE Transactions on Medical Imaging*, vol. 22, pp. 986–1004, Aug 2003.

[6] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "Voxelmorph: A learning framework for deformable medical image registration," *IEEE Transactions on Medical Imaging*, vol. 38, pp. 1788–1800, Aug 2019.

[7] J.-N. Vauthey, E. K. Abdalla, D. A. Doherty, P. Gertsch, M. J. Fenstermacher, E. M. Loyer, J. Lerut, R. Materne, X. Wang, A. Encarnacion, D. Herron, C. Mathey, G. Ferrari, C. Charnsangavej, K.-A. Do, and A. Denys, "Body surface area and body weight predict total liver volume in western adults," *Liver Transplantation*, vol. 8, no. 3, pp. 233–240, 2002.

[8] S. R. Abdel-Misih and M. Bloomston, "Liver anatomy," *Surgical Clinics of North America*, vol. 90, no. 4, pp. 643 – 653, 2010.

[9] M. Schünke, E. Schulte, and U. Schumacher, *Prometheus - Lernatlas der Anatomie: Innere Organe ; 118 Tabellen.* No. Bd. 2 in Prometheus, 2009.

[10] J. K. Suh, J. Lee, J.-H. Lee, S. Shin, H. j. Tchoe, and J.-W. Kwon, "Risk factors for developing liver cancer in people with and without liver disease," *PLOS ONE*, vol. 13, pp. 1–13, 10 2018.

[11] X. Guan, "Cancer metastases: challenges and opportunities," *Acta Pharmaceutica Sinica B*, vol. 5, no. 5, pp. 402 – 418, 2015.

[12] "Metastatic cancer." `https://www.cancer.gov/types/metastatic-cancer?redirect=true`. Accessed: 2019-07-01.

[13] P. Rawla, T. Sunkara, P. Muralidharan, and J. P. Raj, "Update in global trends and aetiology of hepatocellular carcinoma," *Contemporary Oncology/Współczesna Onkologia*, vol. 22, no. 3, pp. 141–150, 2018.

[14] U. P. Neumann, D. Seehofer, and P. Neuhaus, "The Surgical Treatment of Hepatic Metastases in Colorectal Carcinoma," *Deutsches Ärzteblatt International*, vol. 107, no. 19, pp. 335–342, 2010.

[15] J. Heitmann and M. Guckenberger, "Perspectives on oligometastasis: challenges and opportunities," *Journal of Thoracic Disease*, vol. 10, no. 1, 2018.

[16] A. Tannapfel, H.-P. Dienes, and A. W. Lohse, "The Indications for Liver Biopsy," *Deutsches Ärzteblatt International*, vol. 109, no. 27-28, pp. 477–483, 2012.

[17] H. Qiu, A. W. Katz, and M. T. Milano, "Oligometastases to the liver: predicting outcomes based upon radiation sensitivity," *Journal of Thoracic Disease*, vol. 8, no. 10, pp. 113–117, 2016.

[18] C.-Y. Liu, K.-F. Chen, and P.-J. Chen, "Treatment of liver cancer," *Cold Spring Harbor Perspectives in Medicine*, vol. 5, no. 9, 2015.

[19] K. Cleary and T. M. Peters, "Image-guided interventions: Technology review and clinical applications," *Annual Review of Biomedical Engineering*, vol. 12, no. 1, pp. 119–142, 2010.

[20] A. Schenk, D. Haemmerich, and T. Preusser, "Planning of image-guided interventions in the liver," *IEEE Pulse*, vol. 2, pp. 48–55, Sep. 2011.

[21] W. C. RÖNTGEN, "On a new kind of rays," *Science*, vol. 3, no. 59, pp. 227–231, 1896.

[22] V. Petrik, V. Apok, J. A. Britton, B. A. Bell, and M. C. Papadopoulos, "Godfrey Hounsfield and the Dawn of Computed Tomography," *Neurosurgery*, vol. 58, pp. 780–787, 04 2006.

[23] A. Katsevich, "An improved exact filtered backprojection algorithm for spiral computed tomography," *Advances in Applied Mathematics*, vol. 32, no. 4, pp. 681 – 697, 2004.

[24] M. Beister, D. Kolditz, and W. A. Kalender, "Iterative reconstruction methods in x-ray ct," *Physica Medica*, vol. 28, no. 2, pp. 94 – 108, 2012.

[25] L. W. Goldman, "Principles of ct: multislice ct.," *Journal of nuclear medicine technology*, vol. 36 2, pp. 57–68, 2008.

[26] M. Sarti, W. P. Brehmer, and S. B. Gay, "Low-dose techniques in ct-guided interventions," *RadioGraphics*, vol. 32, no. 4, pp. 1109–1119, 2012.

[27] R. C. Orth, M. J. Wallace, and M. D. Kuo, "C-arm cone-beam ct: General principles and technical considerations for use in interventional radiology," *Journal of Vascular and Interventional Radiology*, vol. 20, no. 7, Supplement, pp. S538 – S544, 2009.

[28] R. Schulze, U. Heil, D. Groß, D. Bruellmann, E. Dranischnikow, U. Schwanecke, and E. Schoemer, "Artefacts in CBCT: a review," *Dentomaxillofacial Radiology*, vol. 40, no. 5, pp. 265–273, 2011. PMID: 21697151.

[29] F. Verdun, D. Racine, J. Ott, M. Tapiovaara, P. Toroi, F. Bochud, W. Veldkamp, A. Schegerer, R. Bouwman, I. H. Giron, N. Marshall, and S. Edyvean, "Image quality in ct: From physical measurements to model observers," *Physica Medica*, vol. 31, no. 8, pp. 823 – 843, 2015.

[30] L. Zhu, Y. Xie, J. Wang, and L. Xing, "Scatter correction for cone-beam ct in radiation therapy," *Medical Physics*, vol. 36, pp. 2258–2268, 2009.

[31] J. H. Siewerdsen and D. A. Jaffray, "Cone-beam computed tomography with a flat-panel imager: Magnitude and effects of x-ray scatter," *Medical Physics*, vol. 28, no. 2, pp. 220–231, 2001.

[32] R. Fahrig, R. Dixon, T. Payne, R. L. Morin, A. Ganguly, and N. Strobel, "Dose and image quality for a cone-beam c-arm ct system," *Medical Physics*, vol. 33, no. 12, pp. 4541–4550, 2006.

[33] P. Mah, T. E. Reeves, and W. D. McDavid, "Deriving hounsfield units using grey levels in cone beam computed tomography," *Dentomaxillofacial Radiology*, vol. 39, no. 6, pp. 323–335, 2010. PMID: 20729181.

[34] T. Reeves, P. Mah, and W. McDavid, "Deriving hounsfield units using grey levels in cone beam ct: a clinical application," *Dentomaxillofacial Radiology*, vol. 41, no. 6, pp. 500–508, 2012. PMID: 22752324.

[35] W. G. Bradley, "History of medical imaging," *Proceedings of the American Philosophical Society*, vol. 152, no. 3, pp. 349–361, 2008.

[36] G. H. Glover, "Overview of functional magnetic resonance imaging," *Neurosurgery Clinics of North America*, vol. 22, no. 2, pp. 133 – 139, 2011.

[37] S. Y. Huang, R. T. Seethamraju, P. Patel, P. F. Hahn, J. E. Kirsch, and A. R. Guimaraes, "Body mr imaging: Artifacts, k-space, and solutions," *RadioGraphics*, vol. 35, no. 5, pp. 1439–1460, 2015. PMID: 26207581.

[38] M. N. Hood, V. B. Ho, J. G. Smirniotopoulos, and J. Szumowski, "Chemical shift: The artifact and clinical tool revisited," *RadioGraphics*, vol. 19, no. 2, pp. 357–371, 1999. PMID: 10194784.

[39] A. Sotiras, C. Davatzikos, and N. Paragios, "Deformable medical image registration: A survey," *IEEE Transactions on Medical Imaging*, vol. 32, pp. 1153–1190, July 2013.

[40] B. Zitova and J. Flusser, "Image registration methods: a survey," *Image and Vision Computing*, vol. 21, no. 11, pp. 977 – 1000, 2003.

[41] J. Wang, Y. Liu, J. H. Noble, and B. M. Dawant, "Automatic selection of landmarks in T1-weighted head MRI with regression forests for image registration initialization," in *Medical Imaging 2017: Image Processing*, vol. 10133, pp. 735 – 747, 2017.

[42] M. R. M. Jing Wu, Emam Elhak Abdel-Fatah, "Fully automatic initialization of two-dimensional–three-dimensional medical image registration using hybrid classifier," *Journal of Medical Imaging*, vol. 2, no. 2, pp. 1 – 10, 2015.

[43] J. Rackerseder, M. Baust, R. Göbl, N. Navab, and C. Hennersperger, "Initialize globally before acting locally: Enabling landmark-free 3d us to mri registration," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 827–835, 2018.

[44] K. S. Arun, T. S. Huang, and S. D. Blostein, "Least-squares fitting of two 3-d point sets," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-9, pp. 698–700, Sep. 1987.

[45] D. L. G. Hill, D. J. Hawkes, J. E. Crossman, M. J. Gleeson, T. C. S. Cox, E. E. C. M. L. Bracey, A. J. Strong, and P. Graves, "Registration of mr and ct images for skull base surgery using point-like anatomical features," *The British Journal of Radiology*, vol. 64, no. 767, pp. 1030–1035, 1991. PMID: 1742584.

[46] J. B. A. Maintz and M. A. Viergever, "An overview of medical image registration methods," tech. rep., In Symposium of the Belgian hospital physicists association (SBPH-BVZF, 1996.

[47] L. S. C. Prado, E. A. V. Díaz, and R. Romo, "Elastic models application for thorax image registration," *Journal of Physics: Conference Series*, vol. 90, p. 012055, nov 2007.

[48] J. Kybic and M. Unser, "Fast parametric elastic image registration," *IEEE Transactions on Image Processing*, vol. 12, pp. 1427–1442, Nov 2003.

[49] A. A. Moghe and J. Singhai, "Image registration: A review of elastic registration methods applied to medical imaging," *International Journal of Computer Applications*, vol. 70, no. 7, 2013.

[50] A. Pawar, Y. Zhang, Y. Jia, X. Wei, T. Rabczuk, C. L. Chan, and C. Anitescu, "Adaptive fem-based nonrigid image registration using truncated hierarchical b-splines," *Computers and Mathematics with Applications*, vol. 72, no. 8, pp. 2028 – 2040, 2016.

[51] J. A. Schnabel, C. Tanner, A. D. Castellano-Smith, A. Degenhard, M. O. Leach, D. R. Hose, D. L. G. Hill, and D. J. Hawkes, "Validation of nonrigid image registration using finite-element methods: application to breast mr images," *IEEE Transactions on Medical Imaging*, vol. 22, pp. 238–247, Feb 2003.

[52] H.-H. Chang and C.-Y. Tsai, "Adaptive registration of magnetic resonance images based on a viscous fluid model," *Computer Methods and Programs in Biomedicine*, vol. 117, no. 2, pp. 80 – 91, 2014.

[53] E. D'Agostino, F. Maes, D. Vandermeulen, and P. Suetens, "A viscous fluid model for multimodal non-rigid image registration using mutual information," *Medical Image Analysis*, vol. 7, no. 4, pp. 565 – 575, 2003.

[54] C. Rong, J. Zhou, L. Luo, and G. Cao, "Fast non-rigid image registration using viscous fluid model and b-spline," in *2008 2nd International Conference on Bioinformatics and Biomedical Engineering*, pp. 2450–2453, May 2008.

[55] J.-P. Thirion, "Image matching as a diffusion process: an analogy with maxwell's demons," *Medical Image Analysis*, vol. 2, no. 3, pp. 243 – 260, 1998.

[56] W. Bai and M. Brady, "Regularized b-spline deformable registration for respiratory motion correction in PET images," *Physics in Medicine and Biology*, vol. 54, pp. 2719–2736, apr 2009.

[57] T. J. Jacobson and M. J. Murphy, "Optimized knot placement for b-splines in deformable image registration," *Medical Physics*, vol. 38, no. 8, pp. 4579–4582, 2011.

[58] J. A. Shackleford, N. Kandasamy, and G. C. Sharp, "On developing b-spline registration algorithms for multi-core processors," *Physics in Medicine and Biology*, vol. 55, pp. 6329–6351, oct 2010.

[59] J. Tsao, "Interpolation artifacts in multimodality image registration based on maximization of mutual information," *IEEE Transactions on Medical Imaging*, vol. 22, pp. 854–864, July 2003.

[60] I. Aganj, B. T. T. Yeo, M. R. Sabuncu, and B. Fischl, "On removing interpolation and resampling artifacts in rigid image registration," *IEEE Transactions on Image Processing*, vol. 22, pp. 816–827, Feb 2013.

[61] H. Ólafsdóttir, H. Pedersen, M. S. Hansen, H. Larsson, and R. Larsen, "Improving image registration by correspondence interpolation," in *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 1524–1527, March 2011.

[62] D. Forsberg, *Robust Image Registration for Improved Clinical Efficiency : Using Local Structure Analysis and Model-Based Processing.* PhD thesis, Linköping University, Medical Informatics, The Institute of Technology, Center for Medical Image Science and Visualization (CMIV), 2013.

[63] T. M. Deserno, C. Gönner, and K. Spitzer, "Survey: interpolation methods in medical image processing," *IEEE Transactions on Medical Imaging*, vol. 18, pp. 1049–1075, 1999.

[64] J. N. Sarvaiya, S. Patnaik, and S. Bombaywala, "Image registration by template matching using normalized cross-correlation," in *2009 International Conference on Advances in Computing, Control, and Telecommunication Technologies*, pp. 819–822, Dec 2009.

[65] B. Avants, C. Epstein, M. Grossman, and J. Gee, "Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain," *Medical Image Analysis*, vol. 12, no. 1, pp. 26 – 41, 2008.

[66] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[67] C. Studholme, D. L. G. Hill, and D. J. Hawkes, "Automated 3d registration of truncated mr and ct images of the head," in *Proceedings of the 1995 British Conference on Machine Vision (Vol. 1)*, BMVC '95, pp. 27–36, 1995.

[68] C. Studholme, D. L. G. Hill, and D. J. Hawkes, "Incorporating connected region labelling into automated image registration using mutual information," in *Proceedings of the Workshop on Mathematical Methods in Biomedical Image Analysis*, pp. 23–31, June 1996.

[69] A. Collignon, D. Vandermeulen, P. Suetens, and G. Marchal, "3d multi-modality medical image registration using feature space clustering," in *Computer Vision, Virtual Reality and Robotics in Medicine*, pp. 195–204, 1995.

[70] P. Viola and W. M. Wells, "Alignment by maximization of mutual information," in *Proceedings of IEEE International Conference on Computer Vision*, pp. 16–23, June 1995.

[71] P. Viola and W. M. Wells III, "Alignment by maximization of mutual information," *International Journal of Computer Vision*, vol. 24, pp. 137–154, Sep 1997.

[72] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," *IEEE Transactions on Medical Imaging*, vol. 16, pp. 187–198, April 1997.

[73] D. Mattes, D. R. Haynor, H. Vesselle, T. K. Lewellen, and W. Eubank, "Nonrigid multimodality image registration," in *Medical Imaging: Image Processing*, 2001.

[74] D. Mattes, D. R. Haynor, H. Vesselle, T. K. Lewellen, and W. Eubank, "Pet-ct image registration in the chest using free-form deformations," *IEEE Transactions on Medical Imaging*, vol. 22, pp. 120–128, Jan 2003.

[75] C. Studholme, D. Hill, and D. Hawkes, "An overlap invariant entropy measure of 3d medical image alignment," *Pattern Recognition*, vol. 32, no. 1, pp. 71 – 86, 1999.

[76] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, pp. 91–110, Nov. 2004.

[77] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Comput. Vis. Image Underst.*, vol. 110, pp. 346–359, June 2008.

[78] F. Wan and F. Deng, "An image registration method based on feature matching," in *Advanced Research on Computer Education, Simulation and Modeling*, pp. 91–95, 2011.

[79] P. J. Besl and N. D. McKay, "A method for registration of 3-d shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, pp. 239–256, Feb 1992.

[80] S. Ruder, "An overview of gradient descent optimization algorithms," *ArXiv*, vol. abs/1609.04747, 2016.

[81] G. B. Dantzig, A. Orden, and P. Wolfe, "The generalized simplex method for minimizing a linear form under linear inequality restraints.," *Pacific J. Math.*, vol. 5, no. 2, pp. 183–195, 1955.

[82] H. O. Hartley, "The modified gauss-newton method for the fitting of non-linear regression functions by least squares," *Technometrics*, vol. 3, no. 2, pp. 269–280, 1961.

[83] R. Byrd, P. Lu, J. Nocedal, and C. Zhu, "A limited memory algorithm for bound constrained optimization," *SIAM Journal of Scientific Computing*, vol. 16, pp. 1190–1208, 9 1995.

[84] M. J. D. Powell, "An efficient method for finding the minimum of a function of several variables without calculating derivatives," *The Computer Journal*, vol. 7, no. 2, p. 155, 1964.

[85] J. J. Moré, "The levenberg-marquardt algorithm: Implementation and theory," in *Numerical Analysis*, pp. 105–116, 1978.

[86] S. Klein, M. Staring, and J. P. W. Pluim, "Evaluation of optimization methods for nonrigid medical image registration using mutual information and b-splines," *IEEE Transactions on Image Processing*, vol. 16, pp. 2879–2890, Dec 2007.

[87] D. Rueckert and P. Aljabar, "Nonrigid registration of medical images: Theory, methods, and applications [applications corner]," *IEEE Signal Processing Magazine*, vol. 27, pp. 113–119, July 2010.

[88] A. Maier, C. Syben, T. Lasser, and C. Riess, "A gentle introduction to deep learning in medical image processing," *Zeitschrift für Medizinische Physik*, vol. 29, no. 2, pp. 86 – 101, 2019.

[89] A.-K. Schnurr, K. Chung, T. Russ, L. R. Schad, and F. G. Z"ollner, "Simulation-based deep artifact correction with convolutional neural networks for limited angle artifacts," *Zeitschrift für Medizinische Physik*, vol. 29, no. 2, pp. 150 – 161, 2019.

[90] A. S. Lundervold and A. Lundervold, "An overview of deep learning in medical imaging focusing on mri," *Zeitschrift für Medizinische Physik*, vol. 29, no. 2, pp. 102 – 127, 2019.

[91] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, vol. 15 of *Proceedings of Machine Learning Research*, pp. 315–323, 11–13 Apr 2011.

[92] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun, "What is the best multi-stage architecture for object recognition?," in *2009 IEEE 12th International Conference on Computer Vision*, pp. 2146–2153, Sep. 2009.

[93] M. Lewenstein and A. Nowak, "Fully connected neural networks with self-control of noise levels," *Phys. Rev. Lett.*, vol. 62, pp. 225–228, Jan 1989.

[94] N. Brunel, "Dynamics of sparsely connected networks of excitatory and inhibitory spiking neurons," *Journal of Computational Neuroscience*, vol. 8, pp. 183–208, May 2000.

[95] Y. LeCun, B. E. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. E. Hubbard, and L. D. Jackel, "Handwritten digit recognition with a back-propagation network," in *Advances in Neural Information Processing Systems 2*, pp. 396–404, 1990.

[96] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *Proceedings of the 30th International Conference on Machine Learning*, vol. 28, pp. 1139–1147, 17–19 Jun 2013.

[97] Q. V. Le, J. Ngiam, A. Coates, A. Lahiri, B. Prochnow, and A. Y. Ng, "On optimization methods for deep learning," in *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pp. 265–272, 2011.

[98] Y. LeCun, Fu Jie Huang, and L. Bottou, "Learning methods for generic object recognition with invariance to pose and lighting," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, vol. 2, p. 104, June 2004.

[99] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proceedings of the 26th Annual International Conference on Machine Learning*, p. 609–616, 2009.

[100] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241, 2015.

[101] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60 – 88, 2017.

[102] H. Wang, L. Dong, J. Daniel, R. Mohan, A. S. Garden, K. K. Ang, D. A. Kuban, M. Bonnen, J. Y. Chang, and R. Cheung, "Validation of an accelerated demons algorithm for deformable image registration in radiation therapy," *Physics in Medicine and Biology*, vol. 50, pp. 2887–2905, jun 2005.

[103] M. Urschler, S. Kluckner, and H. Bischof, "A framework for comparison and evaluation of nonlinear intra-subject image registration algorithms," in *Insight Journal*, 2007.

[104] J. A. Schnabel, C. Tanner, A. D. Castellano-Smith, A. Degenhard, M. O. Leach, D. R. Hose, D. L. G. Hill, and D. J. Hawkes, "Validation of nonrigid image registration using finite-element methods: application to breast mr images," *IEEE Transactions on Medical Imaging*, vol. 22, pp. 238–247, Feb 2003.

[105] W. R. Crum, D. Rueckert, M. Jenkinson, D. Kennedy, and S. M. Smith, "A framework for detailed objective comparison of non-rigid registration algorithms in neuroimaging," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2004*, pp. 679–686, 2004.

[106] P. Hellier, C. Barillot, I. Corouge, B. Gibaud, G. Le Goualher, D. L. Collins, A. Evans, G. Malandain, N. Ayache, G. E. Christensen, and H. J. Johnson, "Retrospective evaluation of intersubject brain registration," *IEEE Transactions on Medical Imaging*, vol. 22, pp. 1120–1130, Sep. 2003.

[107] A. Klein, J. Andersson, B. A. Ardekani, J. Ashburner, B. Avants, M.-C. Chiang, G. E. Christensen, D. L. Collins, J. Gee, P. Hellier, J. H. Song, M. Jenkinson, C. Lepage, D. Rueckert, P. Thompson, T. Vercauteren, R. P. Woods, J. J. Mann, and R. V. Parsey, "Evaluation of 14 nonlinear deformation algorithms applied to human brain mri registration," *NeuroImage*, vol. 46, no. 3, pp. 786 – 802, 2009.

[108] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[109] P. Jaccard, "The distribution of the flora in the alpine zone," *New Phytologist*, vol. 11, pp. 37–50, Feb. 1912.

[110] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, "Comparing images using the hausdorff distance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, pp. 850–863, Sep. 1993.

[111] J. H. Song, G. E. Christensen, J. A. Hawley, Y. Wei, and J. G. Kuhl, "Evaluating image registration using nirep," in *Biomedical Image Registration*, pp. 140–150, 2010.

[112] T. Rohlfing, "Image similarity and tissue overlaps as surrogates for image registration accuracy: Widely used but unreliable," *IEEE Transactions on Medical Imaging*, vol. 31, pp. 153–163, Feb 2012.

[113] J. M. Fitzpatrick, J. B. West, and C. R. Maurer, "Predicting error in rigid-body point-based registration," *IEEE Transactions on Medical Imaging*, vol. 17, pp. 694–702, Oct 1998.

[114] C. R. Maurer, J. J. McCrory, and J. M. Fitzpatrick, "Estimation of accuracy in localizing externally attached markers in multimodal volume head images," in *Medical Imaging 1993: Image Processing*, vol. 1898, pp. 43 – 54, International Society for Optics and Photonics, 1993.

[115] J. M. Fitzpatrick, "Fiducial registration error and target registration error are uncorrelated," in *Medical Imaging 2009: Visualization, Image-Guided Procedures, and Modeling*, vol. 7261, pp. 21 – 32, 2009.

[116] A. Danilchenko and J. M. Fitzpatrick, "General approach to first-order error prediction in rigid point registration," *IEEE Transactions on Medical Imaging*, vol. 30, pp. 679–693, March 2011.

[117] R. Castillo, E. Castillo, R. Guerra, V. E. Johnson, T. McPhail, A. K. Garg, and T. Guerrero, "A framework for evaluation of deformable image registration spatial accuracy using large landmark point sets," *Physics in Medicine and Biology*, vol. 54, pp. 1849–1870, mar 2009.

[118] Z. Wu, E. Rietzel, V. Boldea, D. Sarrut, and G. C. Sharp, "Evaluation of deformable registration of patient lung 4dct with subanatomical region segmentations," *Medical Physics*, vol. 35, no. 2, pp. 775–781, 2008.

[119] A. Pevsner, B. Davis, S. Joshi, A. Hertanto, J. Mechalakos, E. Yorke, K. Rosenzweig, S. Nehmeh, Y. E. Erdi, J. L. Humm, S. Larson, C. C. Ling, and G. S. Mageras, "Evaluation of an automated deformable image matching method for quantifying lung motion in respiration-correlated ct images," *Medical Physics*, vol. 33, no. 2, pp. 369–376, 2006.

[120] J. Vandemeulebroucke, D. Sarrut, and P. Clarysse, "The popi model, a pointvalidated pixel-based breathing thorax model," in *XVth International Conference on the Use of Computers in Radiation Therapy (ICCR)*, 2007.

[121] E. Castillo, R. Castillo, J. Martinez, M. Shenoy, and T. Guerrero, "Four-dimensional deformable image registration using trajectory modeling," *Physics in Medicine and Biology*, vol. 55, pp. 305–327, dec 2009.

[122] R. Castillo, E. Castillo, D. Fuentes, M. Ahmad, A. M. Wood, M. S. Ludwig, and T. Guerrero, "A reference dataset for deformable image registration spatial accuracy evaluation using the COPDgene study archive," *Physics in Medicine and Biology*, vol. 58, pp. 2861–2877, apr 2013.

[123] J. West, J. M. Fitzpatrick, M. Y. Wang, B. M. Dawant, Maurer, *et al.*, "Comparison and evaluation of retrospective intermodality brain image registration techniques," *Journal of Computer Assisted Tomography*, vol. 21, pp. 554–566, Aug. 1997.

[124] K. Murphy, B. van Ginneken, S. Klein, M. Staring, B. de Hoop, M. Viergever, and J. Pluim, "Semi-automatic construction of reference standards for evaluation of image registration," *Medical Image Analysis*, vol. 15, no. 1, pp. 71 – 84, 2011.

[125] K. Murphy, B. van Ginneken, J. M. Reinhardt, S. Kabus, K. Ding, *et al.*, "Evaluation of registration methods on thoracic ct: The empire10 challenge," *IEEE Transactions on Medical Imaging*, vol. 30, pp. 1901–1920, Nov 2011.

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

[127] C. P. Lee, Z. Xu, R. P. Burke, R. Baucom, B. K. Poulose, R. G. Abramson, and B. A. Landman, "Evaluation of five image registration tools for abdominal CT: pitfalls and opportunities with soft anatomy," in *Medical Imaging 2015: Image Processing*, vol. 9413, pp. 434 – 440, International Society for Optics and Photonics, 2015.

[128] D. Skerl, B. Likar, and F. Pernus, "A protocol for evaluation of similarity measures for rigid registration," *IEEE Transactions on Medical Imaging*, vol. 25, pp. 779–791, June 2006.

[129] D. Skerl, B. Likar, and F. Pernus, "A protocol for evaluation of similarity measures for non-rigid registration," *Medical Image Analysis*, vol. 12, no. 1, pp. 42 – 54, 2008.

[130] E. B. van de Kraats, G. P. Penney, D. Tomazevic, T. van Walsum, and W. J. Niessen, "Standardized evaluation methodology for 2-d-3-d registration," *IEEE Transactions on Medical Imaging*, vol. 24, pp. 1177–1189, Sep. 2005.

[131] M. Mellor and M. Brady, "Phase mutual information as a similarity measure for registration," *Medical Image Analysis*, vol. 9, no. 4, pp. 330 – 343, 2005.

[132] M. Droske and M. Rumpf, "Multiscale joint segmentation and registration of image morphology," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, pp. 2181–2194, Dec 2007.

[133] J. P. W. Pluim, J. B. A. Maintz, and M. A. Viergever, "Image registration by maximization of combined mutual information and gradient information," *IEEE Transactions on Medical Imaging*, vol. 19, pp. 809–814, Aug 2000.

[134] E. Hodneland, A. Lundervold, J. Rørvik, and A. Z. Munthe-Kaas, "Normalized gradient fields for nonlinear motion correction of dce-mri time series," *Computerized Medical Imaging and Graphics*, vol. 38, no. 3, pp. 202 – 210, 2014.

[135] E. Haber and J. Modersitzki, "Intensity gradient based registration and fusion of multi-modal images," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 726–733, 2006.

[136] C. Wachinger and N. Navab, "Entropy and laplacian images: Structural representations for multi-modal registration," *Medical Image Analysis*, vol. 16, no. 1, pp. 1 – 17, 2012.

[137] E. Shechtman and M. Irani, "Matching local self-similarities across images and videos," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, June 2007.

[138] Jing Huang, Suya You, and Jiaping Zhao, "Multimodal image matching using self similarity," in *2011 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pp. 1–6, Oct 2011.

[139] J. Zhou and Q. Liu, "A combined similarity measure for multimodal image registration," in *2015 IEEE International Conference on Imaging Systems and Techniques (IST)*, pp. 1–5, Sep. 2015.

[140] H. Rivaz, Z. Karimaghaloo, and D. L. Collins, "Self-similarity weighted mutual information: A new nonrigid image registration metric," *Medical Image Analysis*, vol. 18, no. 2, pp. 343 – 358, 2014.

[141] M. P. Heinrich, M. Jenkinson, M. Bhushan, T. Matin, F. V. Gleeson, S. M. Brady, and J. A. Schnabel, "Mind: Modality independent neighbourhood descriptor for multi-modal deformable registration," *Medical Image Analysis*, vol. 16, no. 7, pp. 1423 – 1435, 2012.

[142] D. Jiang, Y. Shi, X. Chen, M. Wang, and Z. Song, "Fast and robust multimodal image registration using a local derivative pattern," *Medical Physics*, vol. 44, no. 2, pp. 497–509, 2017.

[143] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, pp. 886–893 vol. 1, June 2005.

[144] O. Déniz, G. Bueno, J. Salido, and F. D. la Torre, "Face recognition using histograms of oriented gradients," *Pattern Recognition Letters*, vol. 32, pp. 1598 – 1603, 2011.

[145] P. E. Rybski, D. Huber, D. D. Morris, and R. Hoffman, "Visual classification of coarse vehicle orientation using histogram of oriented gradients features," in *IEEE Intelligent Vehicles Symposium*, pp. 921–928, 2010.

[146] A. Misra, T. Abe, and K. Deguchi, "Hand gesture recognition using histogram of oriented gradients and partial least squares regression," in *MVA2011 IAPR Conference on Machine Vision Applications*, pp. 479–482, 2011.

[147] E. Abraham, S. Mishra, N. Tripathi, and G. Sukumaran, "HOG descriptor based registration (a new image registration technique)," in *15th International Conference on Advanced Computing Technologies (ICACT)*, pp. 1–4, 2013.

[148] W. Zhou, L. Zhang, Y. Xie, and C. Liang, "A novel technique for prealignment in multimodality medical image registration," *BioMed Research International*, 2014.

[149] S. Kim, D. Min, B. Ham, S. Lin, and K. Sohn, "Fcss: Fully convolutional self-similarity for dense semantic correspondence," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, pp. 581–595, March 2019.

[150] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "BRIEF: Binary robust independent elementary features," in *European Conference on Computer Vision (ECCV)*, 2010.

[151] X. Cheng, L. Zhang, and Y. Zheng, "Deep similarity learning for multimodal medical images," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 6, no. 3, pp. 248–252, 2018.

[152] M. Simonovsky, B. Gutiérrez-Becker, D. Mateus, N. Navab, and N. Komodakis, "A deep metric for multimodal registration," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 10–18, 2016.

[153] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4353–4361, 2015.

[154] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. Moreno-Noguer, "Discriminative learning of deep convolutional feature point descriptors," in *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 118–126, Dec 2015.

[155] B. G. V. Kumar, G. Carneiro, and I. D. Reid, "Learning local image descriptors with deep siamese and triplet convolutional networks by minimizing global loss functions," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5385–5394, 2016.

[156] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *Similarity-Based Pattern Recognition*, pp. 84–92, 2015.

[157] G. Haskins, U. Kruger, and P. Yan, "Deep learning in medical image registration: a survey," *Machine Vision and Applications*, vol. 31, p. 8, Jan 2020.

[158] M.-M. Rohé, M. Datar, T. Heimann, M. Sermesant, and X. Pennec, "Svf-net: Learning deformable image registration using shape matching," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2017.

[159] J. Krebs, T. Mansi, H. Delingette, L. Zhang, F. C. Ghesu, S. Miao, A. Maier, N. Ayache, R. Liao, and A. Kamen, "Robust non-rigid registration through agent-based action learning," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 344–352, Sept. 2017.

[160] H. Sokooti, B. de Vos, F. Berendsen, B. P. F. Lelieveldt, I. Išgum, and M. Staring, "Nonrigid image registration using multi-scale 3d convolutional neural networks," in *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 232–239, 2017.

[161] X. Yang, R. Kwitt, M. Styner, and M. Niethammer, "Quicksilver: Fast predictive image registration – a deep learning approach," *NeuroImage*, vol. 158, pp. 378 – 396, 2017.

[162] H. Li and Y. Fan, "Non-rigid image registration using fully convolutional networks with deep self-supervision," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI)*, 2018.

[163] B. D. de Vos, F. F. Berendsen, M. A. Viergever, M. Staring, and I. Išgum, "End-to-end unsupervised deformable image registration with a convolutional neural network," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2017.

[164] G. Wu, M. Kim, Q. Wang, B. C. Munsell, and D. Shen, "Scalable high-performance image registration framework by unsupervised deep feature representations learning," *IEEE Transactions on Biomedical Engineering*, vol. 63, pp. 1505–1516, July 2016.

[165] Y. Hu, M. Modat, E. Gibson, W. Li, N. Ghavami, E. Bonmati, G. Wang, S. Bandula, C. M. Moore, M. Emberton, S. Ourselin, J. A. Noble, D. C. Barratt, and T. Vercauteren, "Weakly-supervised convolutional neural networks for multimodal image registration," *Medical Image Analysis*, vol. 49, pp. 1 – 13, 2018.

[166] Y. Hu, M. Modat, E. Gibson, W. Li, N. Ghavami, E. Bonmati, G. Wang, S. Bandula, C. M. Moore, M. Emberton, S. Ourselin, J. A. Noble, D. C. Barratt, and T. Vercauteren, "Weakly-supervised convolutional neural networks for multimodal image registration," *Medical Image Analysis*, vol. 49, pp. 1 – 13, 2018.

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

[168] A. V. Dalca, G. Balakrishnan, J. Guttag, and M. R. Sabuncu, "Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces," *Medical Image Analysis*, vol. 57, pp. 226 – 236, 2019.

[169] A. V. Dalca, G. Balakrishnan, J. Guttag, and M. R. Sabuncu, "Unsupervised learning for fast probabilistic diffeomorphic registration," in *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 729–738, 2018.

[170] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. V. Guttag, and A. V. Dalca, "An unsupervised learning model for deformable medical image registration," *CoRR*, vol. abs/1802.02604, 2018.

[171] M. Jaderberg, K. Simonyan, A. Zisserman, and k. kavukcuoglu, "Spatial transformer networks," in *Advances in Neural Information Processing Systems 28*, pp. 2017–2025, 2015.

[172] L. Ibanez, W. Schroeder, L. Ng, and J. Cates, *The ITK Software Guide*. Kitware, Inc., first ed., 2003. ISBN 1-930934-10-6.

[173] B. Lowekamp, D. Chen, L. Ibanez, and D. Blezek, "The design of simpleitk," *Frontiers in Neuroinformatics*, vol. 7, p. 45, 2013.

[174] K. Marstal, F. Berendsen, M. Staring, and S. Klein, "Simpleelastix: A user-friendly, multilingual library for medical image registration," in *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 574–582, June 2016.

[175] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. W. Pluim, "elastix: A toolbox for intensity-based medical image registration," *IEEE Transactions on Medical Imaging*, vol. 29, pp. 196–205, Jan 2010.

[176] M. Nolden, S. Zelzer, A. Seitel, D. Wald, M. Müller, A. M. Franz, D. Maleike, M. Fangerau, M. Baumhauer, L. Maier-Hein, K. H. Maier-Hein, H. P. Meinzer, and I. Wolf, "The medical imaging interaction toolkit: challenges and advances," *International Journal of Computer Assisted Radiology and Surgery*, vol. 8, no. 4, pp. 607–620, 2013.

[177] W. Schroeder, K. Martin, and B. Lorensen, *The Visualization Toolkit*. Kitware, 2006.

[178] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *CoRR*, vol. abs/1603.04467, 2016.

[179] F. Chollet, "keras." `https://github.com/fchollet/keras`, 2015. Accessed: 2019-08-15.

[180] B. Waldkirch, S. Engelhardt, F. G. Zöllner, L. R. Schad, and I. Wolf, "Multimodal image registration of pre- and intra-interventional data for surgical planning of transarterial chemoembolisation," in *Proc. SPIE, Medical Imaging 2019: Image-Guided Procedures, Robotic Interventions and Modeling*, p. 109512U, 2019.

[181] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2242–2251, Oct 2017.

[182] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit*, pp. 5967–5976, 2017.

[183] J. M. Wolterink, A. M. Dinkla, M. H. F. Savenije, P. R. Seevinck, C. A. T. van den Berg, and I. Išgum, "Deep mr to ct synthesis using unpaired data," in *Simulation and Synthesis in Medical Imaging*, pp. 14–23, 2017.

[184] Y. Hiasa, Y. Otake, M. Takao, T. Matsuoka, K. Takashima, A. Carass, J. L. Prince, N. Sugano, and Y. Sato, "Cross-modality image synthesis from unpaired data using cyclegan," in *Simulation and Synthesis in Medical Imaging*, pp. 31–41, 2018.

[185] H. Yang, J. Sun, A. Carass, C. Zhao, J. Lee, Z. Xu, and J. Prince, "Unpaired brain mr-to-ct synthesis using a structure-constrained cyclegan," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 174–182, 2018.

[186] T. Russ, S. Goerttler, A.-K. Schnurr, D. F. Bauer, S. Hatamikia, L. R. Schad, F. G. Zoellner, and K. Chung, "Synthesis of ct images from digital body phantoms using cyclegan," *International Journal of Computer Assisted Radiology and Surgery*, vol. 14, pp. 1741 – 1750, 2019.

[187] D. F. Bauer, A.-K. Schnurr, T. Russ, S. Goerttler, L. R. Schad, F. G. Zöllner, and K. Chung, "Synthesis of ct images using cyclegans: Enhancement of anatomical accuracy," in *Medical Imaging with Deep Learning*, July 2019.

[188] W. P. Segars, G. Sturgeon, S. Mendonca, J. Grimes, and B. M. W. Tsui, "4d xcat phantom for multimodality imaging research," *Medical Physics*, vol. 37, no. 9, pp. 4902–4915, 2010.

[189] "Ixi dataset." `https://brain-development.org/ixi-dataset/`. Accessed: 2019-07-10.

[190] E. B. Saff and A. B. J. Kuijlaars, "Distributing many points on a sphere," *The Mathematical Intelligencer*, vol. 19, no. 1, pp. 5–11, 1997.

[191] L. D. Stefano, S. Mattoccia, and M. Mola, "An efficient algorithm for exhaustive template matching based on normalized cross correlation," in *12th International Conference on Image Analysis and Processing, 2003.Proceedings.*, pp. 322–327, Sep. 2003.

[192] B. Trimborn, I. Wolf, D. Abu-Sammour, T. Henzler, L. R. Schad, and F. G. Zöllner, "Investigation of 3D histograms of oriented gradients for image-based registration of CT with interventional CBCT," in *Medical Imaging 2017: Image-Guided Procedures, Robotic Interventions, and Modeling*, vol. 10135, pp. 96 – 104, 2017.

[193] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[194] "Voxelmorph source code." `https://github.com/voxelmorph/voxelmorph`. Accessed: 2019-12-18.

[195] G. Kabelitz, B. Trimborn, I. Wolf, L. Schad, and F. Zöllner, "Fast 3d-hog (histogram of oriented gradients) for medical imaging with gpu-support," in *3rd Conference on Image-Guided Interventions & Fokus Neuroradiologie (IGIC)*, p. 9, 2017.

List of Figures

List of Tables

Appendix

This appendix contains additional figures to the results presented in Section 5 and is divided according to its subchapters. The figures presented here are added for the sake of completeness. However, they do not contain additional information to the findings presented in Section 5.

Siamese Deep Metric

Synthetic CBCT/CT Abdominal Data

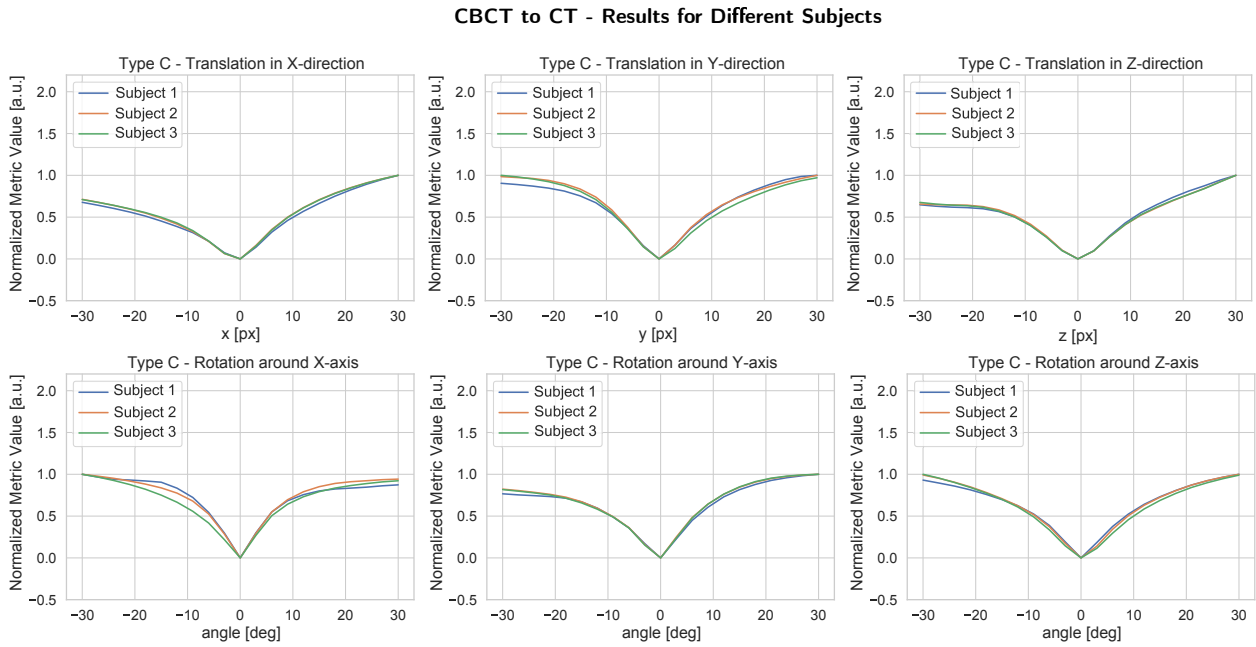


Figure 1: Parametric space obtained for three different pairs of synthetic CBCT and CT data using the SDM trained with patches wit a size of 100 px^3, for translation and rotation transformations.

Synthetic T1/CT Abdominal Data

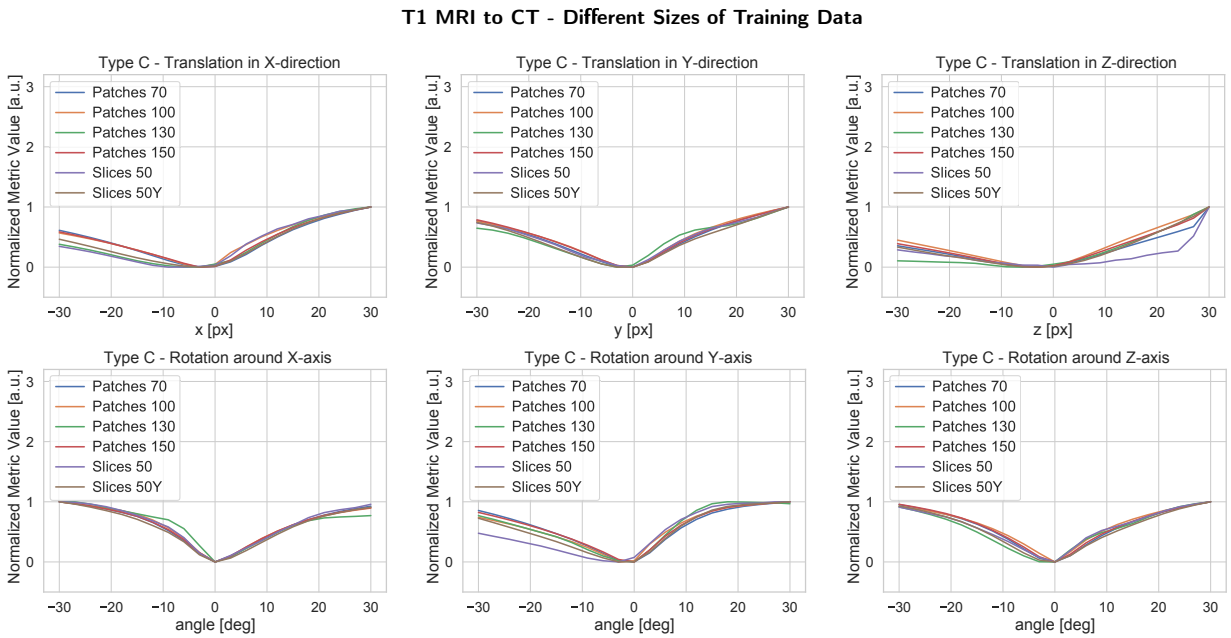


Figure 2: Parametric space obtained for a pair of synthetic T1-MRI and CT data using the SDM trained with different patch sizes, for translation and rotation transformations.

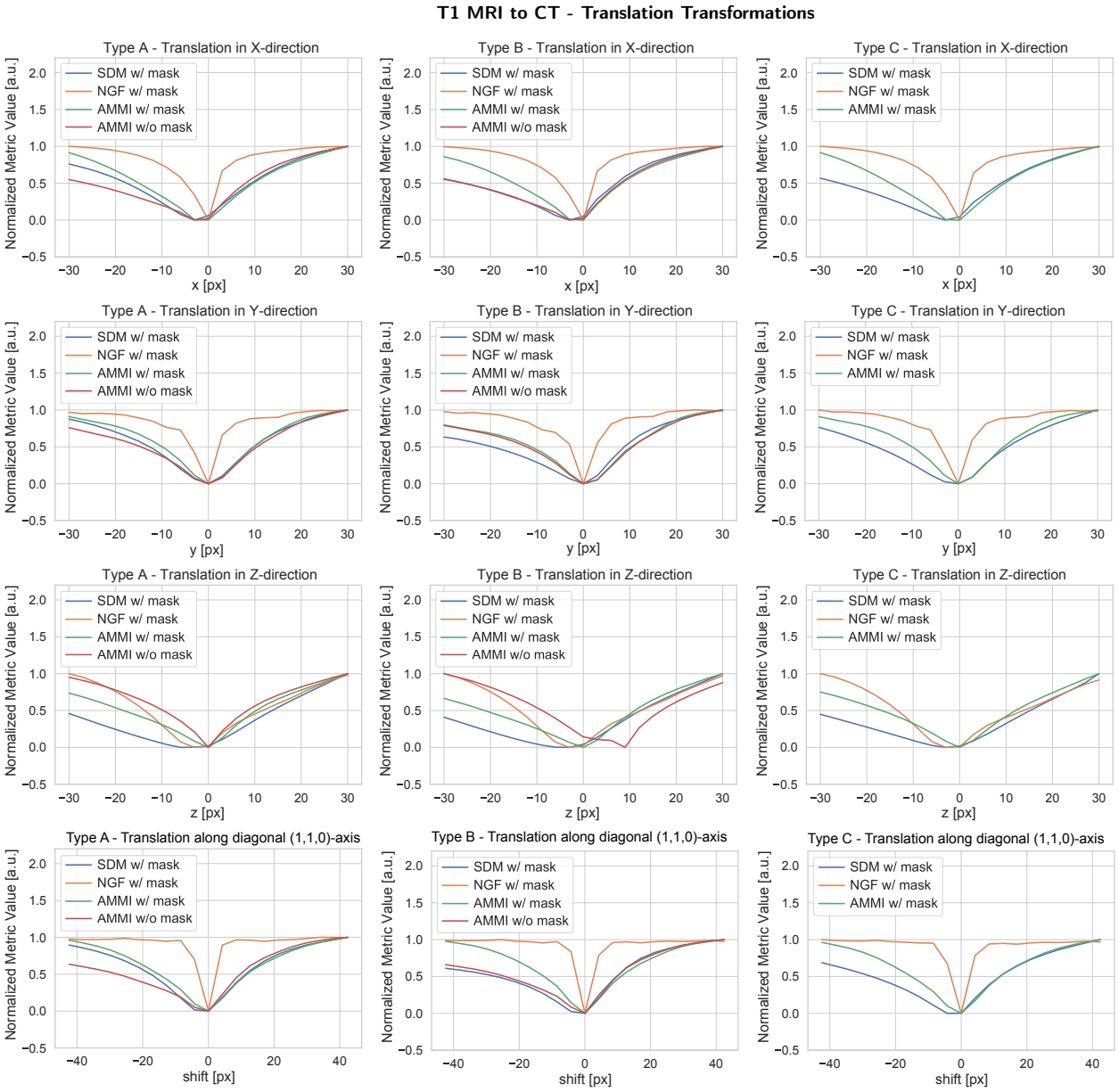


Figure 3: Parametric space obtained for a pair of synthetic T1-MRI and CT data using the learned SDM, a NGF metric and a traditional AMMI metric with and without the use of an image mask for translation in x-,y-,z- and diagonal (1,1,0)-direction. The metric values were projected in a range from 0 - 1

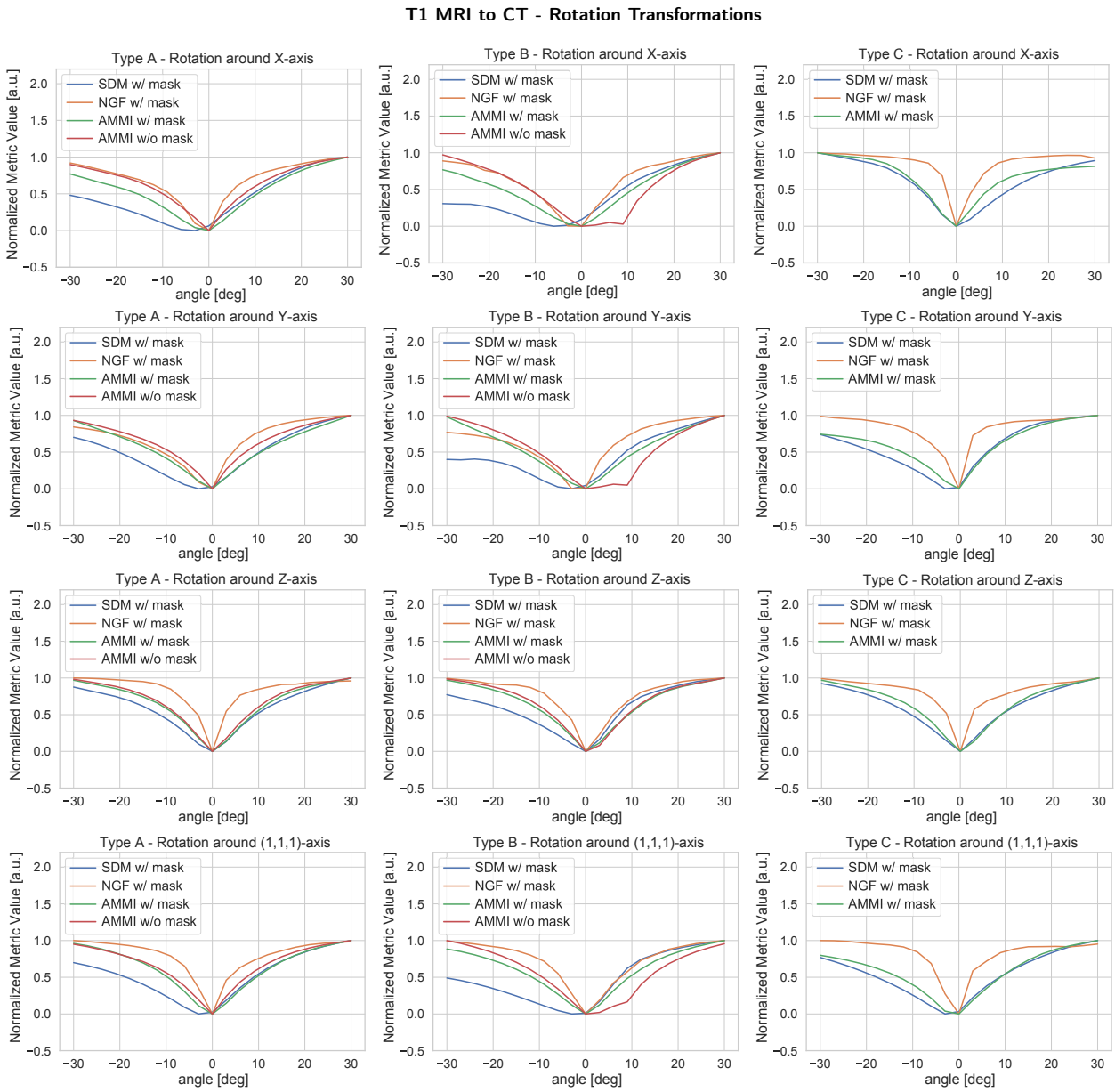


Figure 4: Parametric space obtained for a pair of synthetic T1-MRI and CT data using the learned SDM, a NGF metric and a traditional AMMI metric with and without the use of an image mask for rotation around the x-,y-,z- and diagonal (1,1,1)-axis.

End-to-End Registration Learning

Monomodal Registration Learning

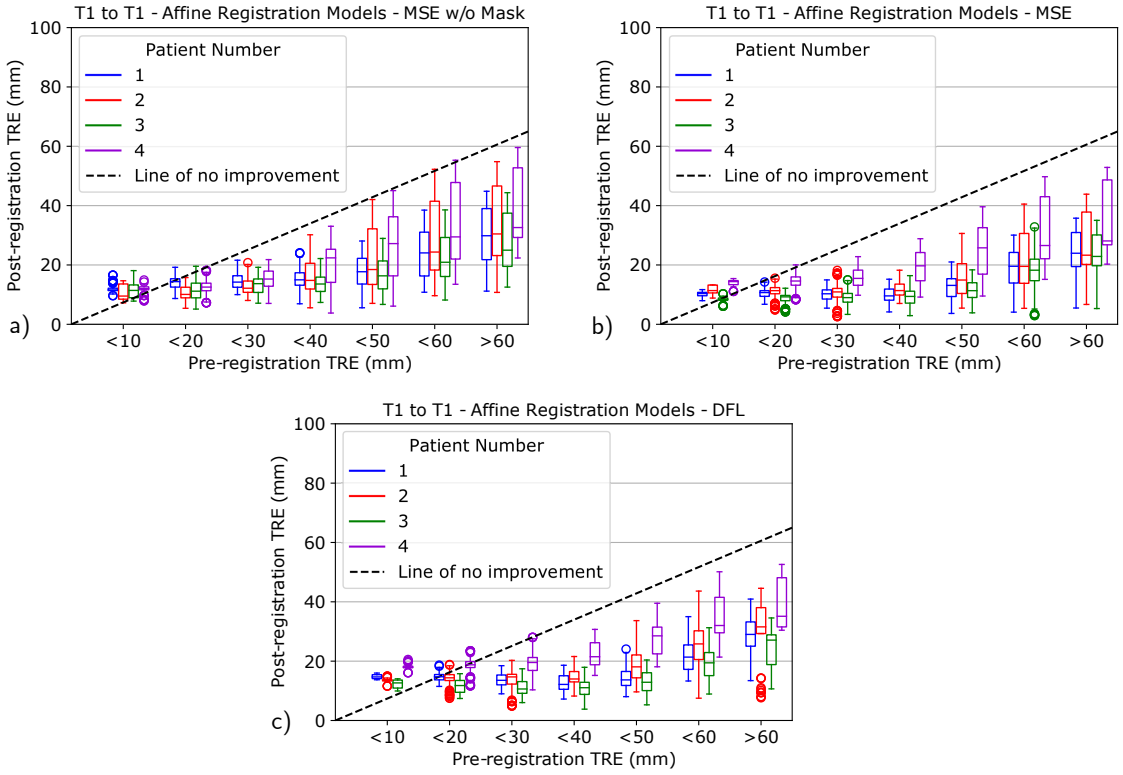


Figure 5: Multimodal affine registration of four different T1 MRI brain data pairs using models trained with different loss functions.

Multimodal Registration Learning

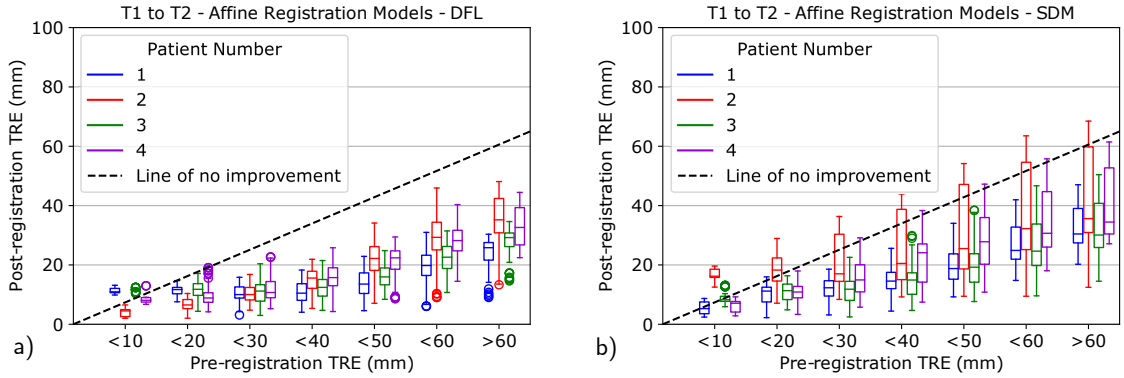


Figure 6: Multimodal affine registration of four different T1 and T2 brain MRI data pairs using models trained with different loss functions.

Transfer to Patient Data

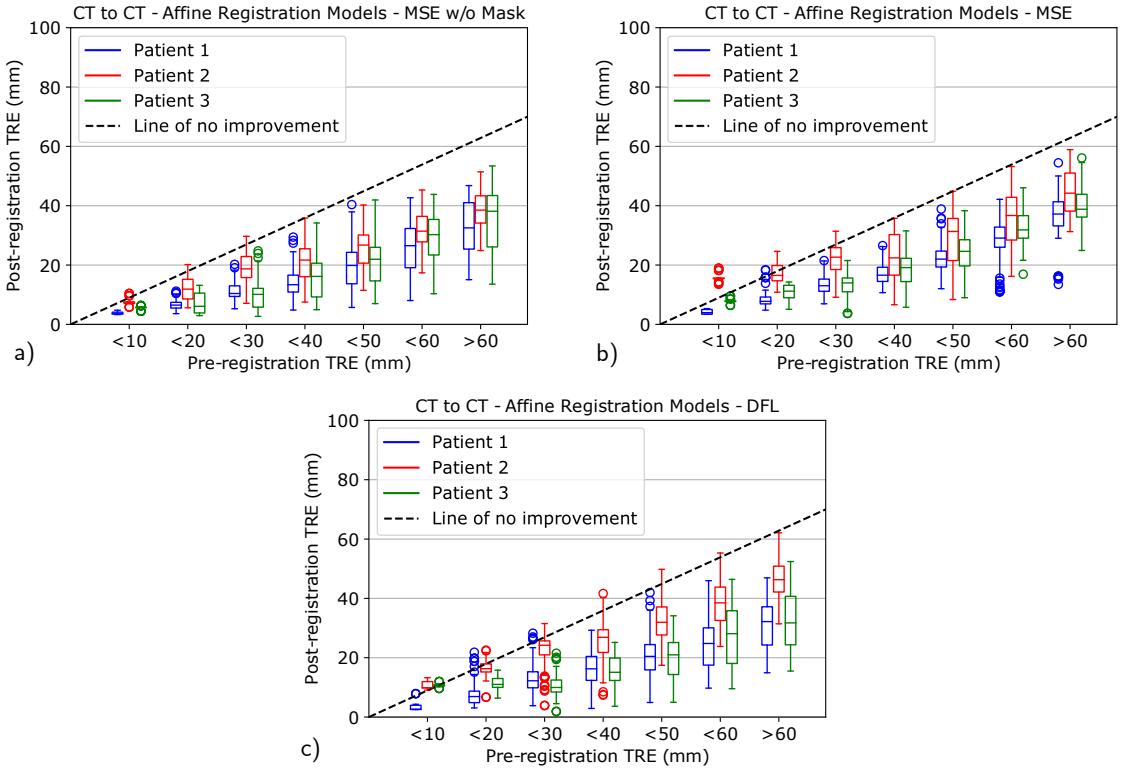


Figure 7: Multimodal affine registration of three different abdominal CT data pairs using models trained with different loss functions.

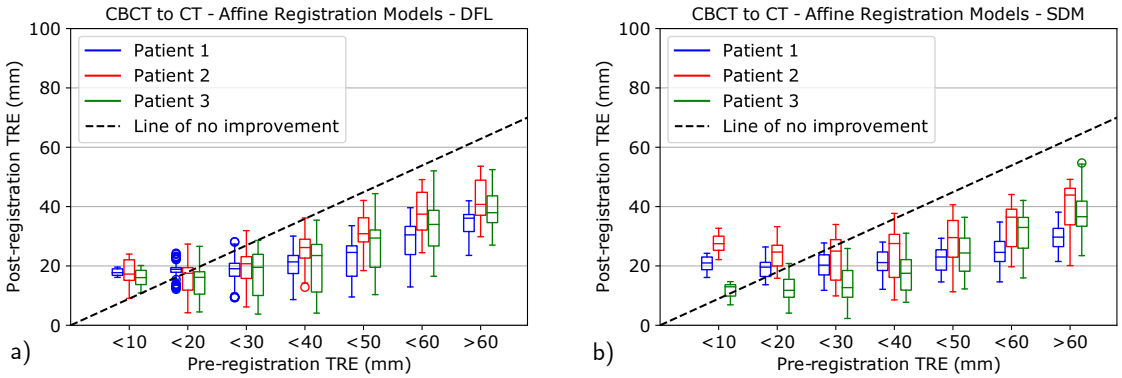


Figure 8: Multimodal affine registration of three different abdominal CBCT/CT data pairs using models trained different loss functions.

List of Publications

Parts of this thesis have been published in the following conference proceedings or abstracts.

Peer-Reviewed Conference Proceedings

- **B. Waldkirch**, S. Engelhardt, F. Zöllner, L. Schad and I. Wolf. Multimodal image registration of pre- and intra-interventional data for surgical planning of transarterial chemoembolisation. Proceedings of SPIE 10951, Medical Imaging 2019: Image-Guided Procedures, Robotic Interventions and Modeling (2019), DOI: `https://doi.org/10.1117/12.2513664`
- **B. Trimborn**, I. Wolf, D. Abu-Sammour, T. Henzler, L. Schad and F. Zöllner. Investigation of 3D histograms of oriented gradients for image-based registration of CT with interventional CBCT. Proceedings of SPIE 101350, Medical Imaging 2017: Image-Guided Procedures, Robotic Interventions and Modeling (2017), DOI: `https://doi.org/10.1117/12.2255601`

Abstracts

- **B. Waldkirch**, D. Bauer, A. Schnurr, S. Engelhardt, F. Zöllner, L. Schad and I. Wolf. Point-based Evaluation of Multimodal Non-Rigid Image Registration of Synthetic Abdominal Data Generated with a Digital Phantom and a Cycle-Consistent Network. 4th Conference on Image-Guided Interventions (IGIC 2019)
- D. Bauer, A. Schnurr, T. Russ, **B. Waldkirch**, L. Schad, F. Zöllner and K. Chung. Synthesis of CBCT Images from Digital Phantoms Using CycleGANs. 4th Conference on Image-Guided Interventions (IGIC 2019)
- **B. Trimborn**, A. Ilina, S. Engelhardt, F. Zöllner, L. Schad and I. Wolf. Image similarity metric evaluation for multimodal registration of the liver. 3rd Conference on Image-Guided Interventions & Fokus Neuroradiologie (IGIC 2017), p. 6
- G. Kabelitz, **B. Trimborn**, I. Wolf, L. Schad and F. Zöllner. Fast 3D-HOG (Histogram of Oriented Gradients) for Medical Imaging with GPU-Support. 3rd Conference on Image-Guided Interventions & Fokus Neuroradiologie (IGIC 2017), p. 9
- **B. Trimborn**, I. Wolf, D. Abu-Sammour , T. Henzler, L. Schad and F. Zöllner. 3D Histograms of Oriented Gradients zur Registrierung von regulären CT mit interventionellen CBCT Daten. Bildverarbeitung für die Medizin 2017, p. 154,
  DOI: `https://doi.org/10.1007/978-3-662-54345-0_37`

Curriculum Vitae

Name:	Barbara Ingeborg Waldkirch (née Trimborn)
Date of birth:	15^{th} September 1989
Place of birth:	Ludwigshafen am Rhein
Nationality:	German

Education

since Nov 2015	**Doctoral Candidate**, Heidelberg University, Medical Faculty Mannheim, Computer-Assisted Clinical Medicine, Mannheim Supervisor: Prof. Dr. rer. nat. Lothar R. Schad
Oct 2012 - Sep 2015	**Master of Science**, Karlsruhe Institute of Technology (KIT), Karlsruhe, grade: 1.0 Major: *Physics*, Minor: *Medical Imaging, Nuclear Medicine* Master Thesis: *Investigations on the origin of the grating interferometric visibility contrast obtained with low-brilliance X-ray sources*, grade: 1.0, Institute for Photon Science and Synchrotron Radiaton, KIT
Sep 2013 - Feb 2014	**Studies Abroad** Université Joseph-Fourier, Grenoble, France Major: *Physics*
Oct 2009 - Sep 2012	**Bachelor of Science**, Karlsruhe Institute of Technology (KIT), Karlsruhe, grade: 2.0 Major: *Physics*, Minor: *Economics* Bachelor Thesis: *Estimation of the dark-field contrast of typical contrast agents in biomedical imaging with grating interferometry*, grade: 1.0, Institute for Photon Science and Synchrotron Radiaton, KIT
2000-2009	**Abitur & french Baccalauréat**, Geschwister-Scholl-Gymnasium, Ludwigshafen am Rhein, grade: 2.0 & Mention: bien

Scholarships

02/2017	DAAD conference travel programme for the *SPIE Medical Imaging 2017*

Research Assistance

Nov 2015 – Mar 2020	Research Assistant, **Institute for Medical Informatics**, University of Applied Sciences Mannheim, Mannheim
Apr 2012 – Aug 2014	Student Assistant, **Institute for Photon Science and Synchrotron Radiation**, Karlsruhe Institute of Technology, Karlsruhe
Sep 2013 – Feb 2014	Student Assistant, **European Synchrotron Radiation Facility**, Grenoble, France

Acknowledgements

This thesis was conducted as a joint research project between the Institute for Medical Informatics of the University of Applied Sciences and the Department of Computer-Assisted Clinical Medicine of the University of Heidelberg within the frame of the research campus M^{2}OLIE. This collaboration offered me a wide range of possibilities and I hereby want to express my gratitude to all parties involved.

First of all, I would like to thank Prof. Dr. Lothar Schad for agreeing to be my supervisor and welcoming me in his research group as an external PhD student.

I am very grateful to Prof. Dr. Ivo Wolf for mentoring this thesis, for always taking time to answer questions and discuss ideas and the exceptional support during the last months of this thesis, I truly appreciate it.

Moreover, I would like to thank Prof. Dr. Frank Zöllner for the support with everything related to M^{2}OLIE, for his feedback and all the helpful discussions.

I had a great time at the department of Computer-Assisted Clinical Medicine and would like to thank the whole group for the exceptionally friendly atmosphere. Special thanks go to my two long-term office mates, Khanlian Chung and Gordian Kabelitz, and the "Haus-8-Team" Alena-Kathrin Schnurr, Tom Russ, Christian Tönnes and Dominik Bauer. Thank you for helping with the data acquisition, the moral support and the enjoyable coffee sessions.

I also would like to thank my "second" group and my colleagues at the University of Applied Sciences: Jun.-Prof. Dr. Sandy Engelhardt, Jonas Cordes, Angelo Torelli, Sven Köhler, Simon Sauerzapf and Lalith Sharan, as well as my students Deniz Tas and Anna Ilina for their contribution to my research. A special 'thank you' goes to Sandy for always finding the time to proofread not only this thesis, but also all other publications, and all the helpful feedback.

I am truly grateful to my family and friends for their steady support throughout my entire life. Foremost, I would like to thank my parents and my siblings who have my back no matter what I do. Another special 'thank you' goes to Jule for always being there for me.

Last and most importantly, I would like to thank my husband Chris who supports me in every way possible and always keeps me grounded.

,