Thesis for the Degree of Doctor of Philosophy

Combining Shape and Learning for Medical Image Analysis

Robust, Scalable and Generalizable Registration and Segmentation

Jennifer Alvén

CHALMERS

Department of Electrical Engineering
Chalmers University of Technology
Göteborg, Sweden 2020

Combining Shape and Learning for Medical Image Analysis
Robust, Scalable and Generalizable Registration and Segmentation
Jennifer Alvén
ISBN 978-91-7905-234-8

Doktorsavhandlingar vid Chalmers tekniska högskola
Ny serie nr 4701
ISSN 0346-718X
Computer Vision and Medical Image Analysis group
Department of Electrical Engineering
Chalmers University of Technology
SE–412 96 Göteborg, Sweden

Cover: Sökaren © Erik Olson / Bildupphovsrätt 2019

Typeset by the author using LaTeX.

Chalmers digitaltryck
Göteborg, Sweden 2020

Combining Shape and Learning for Medical Image Analysis
Robust, Scalable and Generalizable Registration and Segmentation
Jennifer Alvén
Department of Electrical Engineering
Chalmers University of Technology

Abstract

Automatic methods with the ability to make accurate, fast and robust assessments of medical images are highly requested in medical research and clinical care. Excellent automatic algorithms are characterized by speed, allowing for scalability, and an accuracy comparable to an expert radiologist. They should produce morphologically and physiologically plausible results while generalizing well to unseen and rare anatomies. Still, there are few, if any, applications where today's automatic methods succeed to meet these requirements.

The focus of this thesis is two tasks essential for enabling automatic medical image assessment, *medical image segmentation* and *medical image registration*. Medical image registration, *i.e.* aligning two separate medical images, is used as an important sub-routine in many image analysis tools as well as in image fusion, disease progress tracking and population statistics. Medical image segmentation, *i.e.* delineating anatomically or physiologically meaningful boundaries, is used for both diagnostic and visualization purposes in a wide range of applications, *e.g.* in computer-aided diagnosis and surgery.

The thesis comprises five papers addressing medical image registration and/or segmentation for a diverse set of applications and modalities, *i.e.* pericardium segmentation in cardiac CTA, brain region parcellation in MRI, multi-organ segmentation in CT, heart ventricle segmentation in cardiac ultrasound and tau PET registration. The five papers propose competitive registration and segmentation methods enabled by machine learning techniques, *e.g.* random decision forests and convolutional neural networks, as well as by shape modelling, *e.g.* multi-atlas segmentation and conditional random fields.

Keywords: Medical image segmentation, medical image registration, machine learning, shape models, multi-atlas segmentation, feature-based registration, convolutional neural networks, random decision forests, conditional random fields.

Acknowledgements

First and foremost, I would like to offer my special thanks to my supervisor Fredrik Kahl. Thank you for sharing interesting and novel ideas, for encouraging autonomy and ambition and for the helpful guidance through the academic jungle. I would also like to express my great appreciation to my co-supervisor Olof Enqvist. Thank you for sharing reassuring wisdom as well as code snippets in time of need.

Further, I wish to acknowledge:

Current and former roommates, Eva Lendaro, Mikaela Åhlén, Fatemeh Shokrollahi Yancheshmeh, Bushra Riaz and Frida Fejne. Thanks for the company and the never-ending patience.

Current and former doctoral students at the department of Electrical Engineering, Carl Toft, Erik Stenborg, Anders Karlsson, Eskil Jörgensen, Samuel Scheidegger, Jonathan Lock, José Iglesias, Lucas Brynte, and others. Thanks for sharing laughter as well as frustration. I would especially like to express my gratitude to Måns Larsson, thanks for always being willing to help and for sharing the PhD struggles over the years.

The WiSE team, Sabine Reinfeldt, Hana Dobsicek Trefna, Eva Lendaro, Silvia Muceli, Helene Lindström and Yvonne Jonsson. Thanks for being great female role models.

All medical research partners. I would especially like to acknowledge Göran Bergström, David Molnar and Ola Hjelmgren as well as Michael Schöll and Kerstin Heurling. Thanks for time and effort spent on producing high-quality medical data.

Co-authors and collaborators, current and former members of the Computer Vision and Medical Image Analysis group, fellow researchers and employees at the department of Electrical Engineering and MedTech West as well as former students at Chalmers University of Technology.

Finally, I would like to express my deepest gratitude to Jonas Ingesson, my former teacher in mathematics, to my loving husband Daniel Gustafsson and to family and friends - none mentioned, none forgotten.

Publications

Included publications

Paper I Jennifer Alvén, Kerstin Heurling, Ruben Smith, Olof Strandberg, Michael Schöll, Oskar Hansson and Fredrik Kahl. "A Deep Learning Approach to MR-less Spatial Normalization for Tau PET Images". *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI),* 355-363, 2019.

Paper II Jennifer Alvén, Fredrik Kahl, Matilda Landgren, Viktor Larsson, Johannes Ulén and Olof Enqvist. "Shape-Aware Label Fusion for Multi-Atlas Frameworks". *Pattern Recognition Letters,* 124:109-117, 2019. Extended version of paper (c).

Paper III Måns Larsson, Jennifer Alvén, and Fredrik Kahl. "Max-margin learning of deep structured models for semantic segmentation". *Scandinavian Conference on Image Analysis (SCIA),* 28–40, 2017.

Paper IV Alexander Norlén, Jennifer Alvén, David Molnar, Olof Enqvist, Rauni Rossi Norrlund, John Brandberg, Göran Bergström and Fredrik Kahl. "Automatic Pericardium Segmentation and Quantification of Epicardial Fat from Computed Tomography Angiography". *Journal of Medical Imaging,* 3(3), 2016.

Paper V Jennifer Alvén, Alexander Norlén, Olof Enqvist and Fredrik Kahl. "Überatlas: Fast and Robust Registration for Multi-atlas Segmentation". *Pattern Recognition Letters,* 80:245–255, 2016. Extended version of paper (a).

Subsidiary publications

(a) Jennifer Alvén, Alexander Norlén, Olof Enqvist and Fredrik Kahl. "Überatlas: Robust Speed-Up of Feature-Based Registration and Multi-Atlas Segmentation". *Scandinavian Conference on Image Analysis (SCIA),* 92–102, 2015. Received the "Best Student Paper Award" at SCIA 2015.

(b) Fredrik Kahl, Jennifer Alvén, Olof Enqvist, Frida Fejne, Johannes Ulén, Johan Fredriksson, Matilda Landgren and Viktor Larsson. "Good Features for Reliable Registration in Multi-Atlas Segmentation". *VISCERAL Challenge at the International Symposium on Biomedical Imaging (ISBI),* 12–17, 2015.

(c) Jennifer Alvén, Fredrik Kahl, Matilda Landgren, Viktor Larsson and Johannes Ulén. "Shape-Aware Multi-Atlas Segmentation". *International Conference on Pattern Recognition (ICPR),* 1101–1106, 2016. Received the "IBM Best Student Paper Award (Track: Biomedical Image Analysis and Applications)" at ICPR 2016.

(d) Frida Fejne, Matilda Landgren, Jennifer Alvén, Johannes Ulén, Johan Fredriksson, Viktor Larsson and Fredrik Kahl. "Multi-atlas Segmentation Using Robust Feature-Based Registration". In *Cloud-Based Benchmarking of Medical Image Analysis,* Springer International Publishing, 203–218, 2017. Extended version of paper (b).

Abbreviations

Methods, models and metrics

ADMM	**A**lternating **D**irection **M**ethod of **M**ultipliers
ANTs	**A**dvanced **N**ormalization **T**ool**s**
CNN	**C**onvolutional **N**eural **N**etwork
CRF	**C**onditional **R**andom **F**ield
DRAMMS	**D**eformable **R**egistration via **A**ttribute **M**atching and **M**utual-**S**aliency weighting
GAN	**G**enerative **A**dversarial **N**etwork
ICP	**I**terative **C**losest **P**oint
IRLS	**I**teratively **R**eweighted **L**east **S**quares
MAPER	**M**ulti-**A**tlas **P**ropagation with **E**nhanced **R**egistration
MRF	**M**arkov **R**andom **F**ield
(N)MI	(**N**ormalized) **M**utual **I**nformation
PCA	**P**rincipal **C**omponent **A**nalysis
RANSAC	**RA**ndom **SA**mple **C**onsensus
ReLU	**Re**ctified **L**inear **U**nit
SAD	**S**um of **A**bsolute **D**istances
SIFT	**S**cale-**I**nvariant **F**eature **T**ransform
SIMPLE	**S**elective and **I**terative **M**ethod for **P**erformance **L**evel **E**stimation
SPM	**S**tatistical **P**arametric **M**apping
SSD	**S**um of **S**quared **D**istances
STAPLE	**S**imultaneous **T**ruth **A**nd **P**erformance **L**evel **E**stimation
SURF	**S**peeded **U**p **R**obust **F**eatures
SVM	**S**upport **V**ector **M**achine
TPS	**T**hin **P**late **S**plines

Medical, modalities and data

AD	**A**lzheimer's **D**isease
ADNI	**A**lzheimer's **D**isease **N**euroimaging **I**nitiative
BMI	**B**ody **M**ass **I**ndex
CAD	**C**omputer-**A**ided **D**iagnosis
CAS	**C**omputer-**A**ssisted **S**urgery
CT(A)	**C**omputed **T**omography (**A**ngiography)
EF(V)	**E**picardial **F**at (**V**olume)
HU	**H**ounsfield **U**nits
MR(I)	**M**agnetic **R**esonance (**I**maging)
PET	**P**ositron **E**mission **T**omography
SCAPIS	**S**wedish **CA**rdio**P**ulmonary bio**I**mage **S**tudy
SUV(R)	**S**tandardized **U**ptake **V**alue (**R**atio)
VISCERAL	**VIS**ual **C**oncept **E**xtraction challenge in **RA**dio**L**ogy

Contents

II Included Publications

Part I

Introductory Chapters

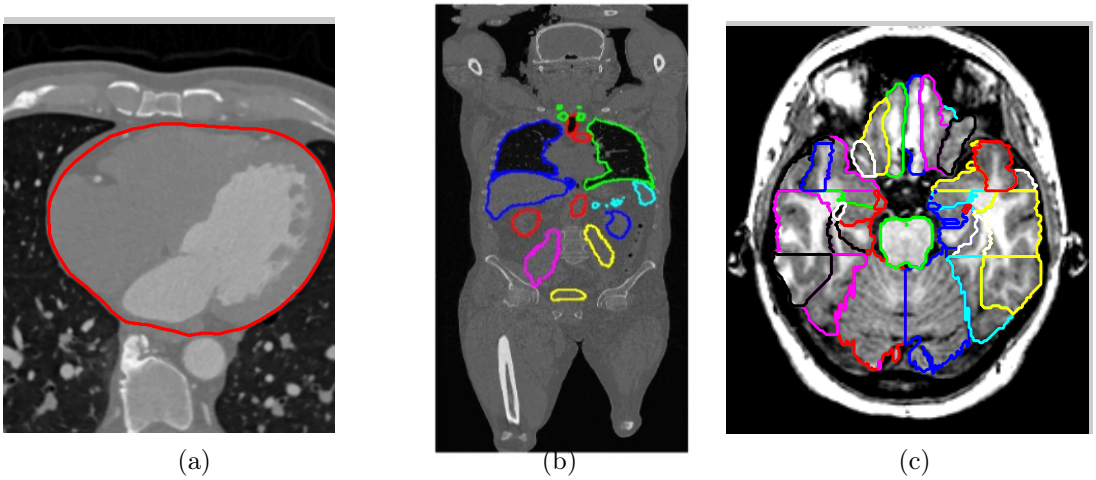
Chapter 1

Introduction

Medical imaging, that is, tools for producing visual representations of the interior (human) body, allows scientists and clinicians to examine, diagnose and treat diseases with means of non-invasive radiology. Medical images, acquired with techniques such as ultrasound, magnetic resonance (MR) imaging, positron emission tomography (PET) and non-enhanced/enhanced computed tomography (CT/CTA), provide information essential for understanding and modeling healthy as well as diseased anatomy and physiology. Decades of successful development of imaging techniques have brought an increased image quality capturing fine anatomical and functional details while the amount of images acquired on a daily basis is steadily growing. The demand for automatic tools for analysis has increased along this development, since manual techniques for inspection cannot effectively and accurately process the huge amount of image data [1].

The field of medical image analysis aims to develop automatic solutions to problems pertaining to medical images. This thesis focuses on two fundamental categories of tasks in this area of research, *medical image segmentation* and *medical image registration*. Automatic segmentation and registration are useful for a wide spectrum of clinical applications, such as computer-aided diagnosis (CAD) systems, treatment planning and in computer-assisted surgery (CAS), including surgery planning, virtual surgery simulation, intra-surgery navigation and robotic surgery, as well as for medical research [2].

Medical image segmentation, the task of dividing an image into meaningful parts by assigning each pixel a label, is an essential problem in medical image analysis and thus utterly well-studied. Commonly, the labels are predetermined and correspond to biologically meaningful object classes, such as different organs or tissue types. The set of labels might correspond to anatomically derived objects embedded in a "background" (for example different organs in whole-body CT), or physiologically derived sub-regions densely covering large parts of the image (for example region parcellation in brain MR image). See Figure 1.1 for three examples of medical segmentation problems. Medical image segmentation has numerous



(a) (b) (c)

Figure 1.1: *Slices of medical 3D images and manual labellings (coloured contours) from three different datasets considered in the included thesis papers. (a) Slice of a* Scapis *[4] cardiac CTA image plus pericardium ("heart sack") labelling. (b) Slice of a* Visceral *[5] whole-body CT image plus organ labellings, such as lungs, liver, kidneys* etc. *(c) Slice of a* Hammers *[6, 7] brain MR image plus region labellings, such as hippocampus, amygdala* etc.

applications. Delineated organ and tissue boundaries are used for both diagnostic and visualization purposes. Examples of tasks are localization of tumors and other pathologies, organ or tissue volume quantification and radiotherapy planning [3].

Medical image registration, the task of establishing spatial correspondences between two separate medical images, is one of the main challenges in contemporary medical image analysis. The images to be registered are typically acquired at different times, with different modalities (medical imaging techniques) or from different subjects. See Figure 1.2 for an example of two aligned cardiac CTA images. Medical image registration is an important pre-processing step in many medical image analysis routines, for instance in segmentation methods. However, the task is also important in itself. One such example is (multi-modal) image fusion, where image registration helps combining images from different modalities or protocols, which facilitates visual comparison in for example CAD and treatment planning. Other applications are monitoring of anatomical or physiological changes over time, including disease progress and growth of pathologies, as well as statistical modeling of population variability and pixelwise comparisons between subjects [8].

Manual registration and segmentation is time-consuming and the quality is highly determined by the expert's skill set. Further, the interobserver variability is usually high. Thus, manual annotation of images is not feasible for applications such as large-scale studies or computer-assisted surgery. Compared to man-

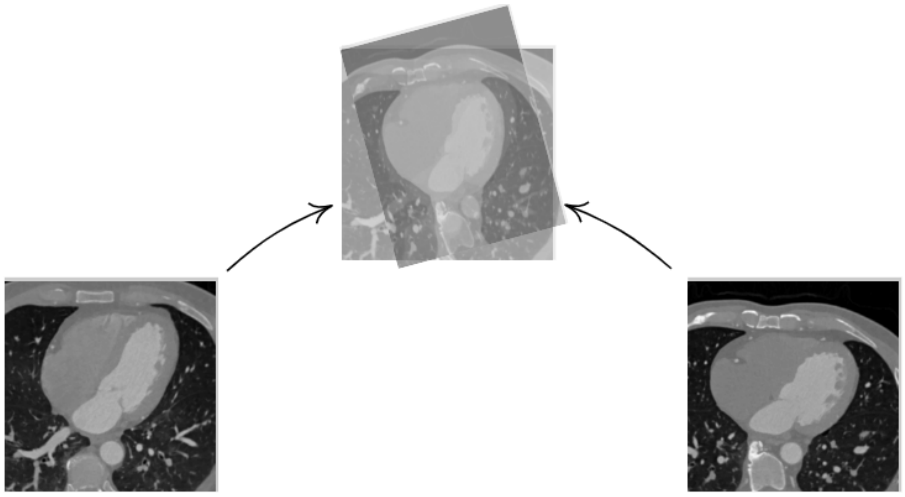


Figure 1.2: *Slices of* Scapis *[4] cardiac CTA images from two different subjects aligned with each other.*

ual methods, automatic segmentation and registration methods are typically fast, cheap, objective and scale well. Accurate automatic methods are therefore highly requested in medical research and by clinical care [9, 10].

Medical images offer several challenges compared to their non-medical counterparts. Typically, medical images contain both low contrast details as well as a moderate to a high level of noise. Inter- and intra-patient variability and imaging ambiguities such as motion artifacts and partial volume effects further increase the difficulty. Compared to neighbouring research fields, such as image analysis and computer vision, manually labelled data is rarely abundant. However, common challenges associated with 2D images, such as (partial) occlusion and light source ambiguities, are usually avoided when processing medical images. Due to these distinct differences (comparing medical images to natural 2D images), the research field includes several analysis methods specifically adapted for medical imaging [3].

1.1 Thesis aim and scope

The included thesis papers propose medical image segmentation and registration methods for several different medical applications. Method development is made with regard to the requirements posed by computer-aided diagnosis and surgery as well as large-scale studies, that is, with respect to (i) accuracy and anatomical/physiological plausibility, (ii) speed and scalability, and (iii) robustness and generalizability.

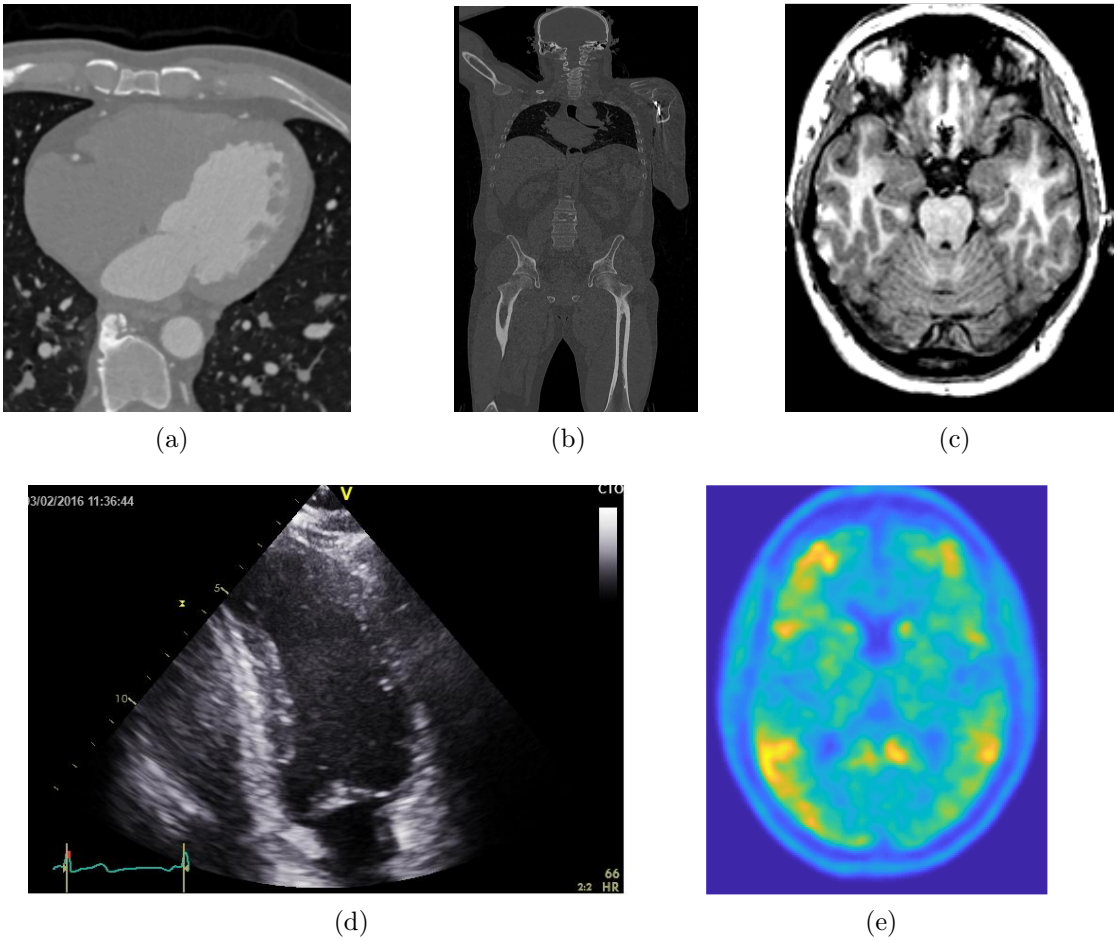
Methods and contributions. Machine learning techniques, *e.g.* random decision forests and convolutional neural networks, are used to construct fast, accurate and robust methods, while shape modelling, *e.g.* multi-atlas segmentation and conditional random fields, provides regularization and ensures plausible results. Combinations of shape and learning are addressed in several of the included publications, as well as in the concluding discussion regarding future research directions.

Paper II-IV focus on developing accurate and robust segmentation methods. Paper II and IV propose two versions of a segmentation pipeline using a combination of multi-atlas segmentation, random decision forests and conditional random field models. In addition, paper II proposes an alternative segmentation pipeline combining multi-atlas segmentation with convolutional neural networks. Paper IV focuses on efficient use of the limited training set by incorporating a generalized formulation of multi-atlas segmentation into the random forest classification framework, while paper II focuses on the qualitative segmentation shape by incorporating an explicit shape prior into the multi-atlas segmentation framework. Paper III also addresses the qualitative segmentation shape, and proposes a segmentation method pairing a convolutional neural network with a conditional random field model that is trainable end-to-end.

Paper I and V proposes two different alternatives to intensity-based image registration. Paper I proposes a deep model including a convolutional neural network regressor as well as differentiable warping, while paper V proposes feature-based image registration including clustering and robust optimization. Both papers focus on increasing the speed, accuracy and generalizability compared to the intensity-based baselines.

Scope and limitations. Typically, medical image analysis methods greatly depend on modality and application, leading to task-specific methods of little use for dissimilar tasks. In this thesis, the proposed methods aim to achieve the opposite, that is, generalizing well across a diverse set of applications and imaging techniques. The included papers consider five significantly different datasets, see Table 1.1 and Figure 1.3. Some of these datasets include very few labelled images. Thus, the included thesis papers must address the shortage of labelled training data when developing and evaluating the proposed methods.

The included papers do not intend to present complete solutions to the registration or segmentation problem at hand, but rather improvements to some parts of the full framework. The included papers do not focus on the technical details for acquiring, pre-processing and annotating medical images. The methods are implemented for research settings, that is, there are no software solutions feasible for everyday use in, for example, clinical care. Finally, the proposed methods should be evaluated on larger datasets before being used in practice.



(a) (b) (c)

(d) (e)

Figure 1.3: *Slices of medical images from five of the datasets considered in the included thesis papers. (a) Slice of a* Scapis *[4] cardiac CTA image. (b) Slice of a* Visceral *[5] whole-body CT image. (c) Slice of a* Hammers *[6, 7] brain MRI. (d) Slice of an* Echo *(in-house) cardiac ultrasound time series. (e) Slice of a* BioFinder *[11] brain tau PET image.*

Table 1.1: *Summary of the datasets included in the thesis publications.*

Name	Modality	Task	Papers
Scapis [4]	cardiac CTA	pericardium segmentation	III, IV, V
Visceral [5]	whole-body CT	multi-organ segmentation	II
Hammers [6, 7]	brain MRI	brain region parcellation	II, V
Echo (in-house)	cardiac ultrasound	heart ventricle segmentation	III
BioFinder [11], Adni *	brain tau PET	multi-modal registration	I

*Alzheimer's Disease Neuroimaging Initiative, https://adni.loni.usc.edu.

1.2 Thesis outline

The thesis is divided into two parts. Part I constitutes the introductory chapters: Chapter 2 briefly compiles theory and methods necessary for understanding the remainder of the thesis, Chapter 3 summarizes the main contributions for each of the included thesis papers and Chapter 4 provides a concluding discussion and potential future research directions. Part II comprises the five included thesis papers.

Chapter 2

Preliminaries

The following sections briefly compile theory, concepts, methods and tools made use of in the included thesis papers and can with ease be skipped by experienced readers. Section 2.1 presents medical images as a concept and lists some common medial imaging techniques. Section 2.2 formalizes the problem of medical image registration and summarizes some common registration methods. Medical image segmentation and two types of commonly used segmentation methods, multi-atlas segmentation and conditional random fields, are accounted for in Section 2.3. Finally, brief introductions to two machine learning tools, random decision forests and convolutional neural networks, are given in Section 2.4.

2.1 Medical images

In this thesis, an image refers to a 2D or 3D matrix whose elements contain intensity levels measured by a medical imaging instrument. A matrix element in a 2D image is referred to as a pixel, while a matrix element in a volumetric image can be referred to as a voxel (VOlume piXEL). In this chapter, the term pixel will be used for both 2D and 3D. The type of imaging technique, *i.e.* type of scanner or probe, that has been used to acquire a medical image is referred to as the modality. The included papers comprise five different modalities, listed below.

Computed Tomography (CT): A CT image is a 3D image produced by a rotating x-ray tube. The 3D image is constructed using measurements of the transmitted x-rays from different angles. CT mainly visualizes morphology and is used for diagnosis of a wide spectrum of diseases, such as bone trauma, abdominal diseases, lung tissue pathology and anatomical changes in the head.

CT angiography (CTA): A CTA image is a CT image where contrast liquid have been injected to the blood vessels. CTA visualizes arteries and veins such as coronary arteries and brain vessels.

Magnetic Resonance (MR) Imaging: A MR image is a 3D image produced by a magnetic field. The 3D image is constructed using measurements from radio frequency signals emitted by excited hydrogen. MR can be used to visualize morphology as well as physiology, and has a wide range of applications, including neuroimaging, cardiovascular imaging and musculoskeletal imaging.

Medical ultrasound: Medical ultrasound (sonography, ultrasonography) uses pulses of ultrasound transmitted from a probe to create 2D or 3D images of the internal body. Medical ultrasound visualizes both anatomy and physiology and is commonly used in obstretics and cardiology.

Positron emission tomography (PET): PET is a nuclear functional imaging technique used to detect molecules in the body. In PET imaging, the scanner detects gamma rays transmitted by positron-emitting radioligands introduced into the body by radioactive tracers. Depending on the radioligand, PET can be used to image, for example, metabolic activity in cancer metastases and amyloid-beta plaques in the brain.

See [12] for a more detailed description of medical imaging and different modalities.

2.2 Medical image registration

To register two images means computing a transformation that aligns one of the images, the source image (the moving image), to the other image, the target image (the fixed/reference image). Image registration algorithms align the source image, $\mathcal{I}_s$, to the target image, $\mathcal{I}_t$, by solving an optimization problem of the form

$$\mathbf{T}^* = \arg\min_{\mathbf{T}} \left[\rho_1(\mathcal{I}_t, \mathbf{T} \circ \mathcal{I}_s) + \rho_2(\mathbf{T})\right], \tag{2.1}$$

where $\mathbf{T}$ is a coordinate transformation from source image pixels to target image pixels and $\mathbf{T} \circ \mathcal{I}_s$ means mapping the source image pixels to the target image space. The level of alignment of the target image and the warped source image is quantified by the first term, ρ_1, while the second term, ρ_2, aims to regularize the transformation, by penalizing implausible deformations and/or by introducing prior knowledge of the deformation. The form of the regularization term should be influenced by the choice of transformation.

Thus, image registration allows for several design choices; type of (i) transformation, (ii) objective function and (iii) optimization method. For a comprehensive overview of different medical image registrations methods and their design choices, see the surveys in [8, 13].

2.2.1 Transformation types

Preferably, the type of transformation is determined by the application. In medical applications, the images are typically first aligned using a rigid and/or an affine transformation followed by a nonlinear local deformation.

The rigid transformation translates, rotates and/or reflects the image globally. Mathematically, it can be described as a composition of an orthogonal map R and a translation $\mathbf{t}$:

$$\mathbf{T}(\mathbf{x}) = R\mathbf{x} + \mathbf{t}, \tag{2.2}$$

where $\mathbf{x}$ is the pixel coordinates.

The affine transformation translates, rotates, scales, reflects and/or shears the image globally. Mathematically, it can be described as a composition of a linear map A and a translation $\mathbf{t}$:

$$\mathbf{T}(\mathbf{x}) = A\mathbf{x} + \mathbf{t}. \tag{2.3}$$

To capture the local nonlinear deformations commonly present in medical applications, the linear transformation is sometimes followed by a non-rigid registration using a nonlinear dense transformation. This deformation is elastic and warps the image locally by using a displacement field $\mathbf{U}$ (that varies with pixels):

$$\mathbf{T}(\mathbf{x}) = \mathbf{x} + \mathbf{U}(\mathbf{x}). \tag{2.4}$$

However, estimating an accurate non-rigid transformation tend to be more computationally demanding than the linear counterpart. Thus, non-rigid registration may be omitted in applications such as computer-assisted surgery or large-scale studies due to timing issues.

2.2.2 Objective functions and optimization methods

The choice of objective function and optimization method is highly influenced by the image registration approach. Roughly speaking, there are two different approaches to image registration; intensity-based registration and feature-based registration. Of course, there are hybrid methods combining advantages of both approaches such as Dramms [14] (Deformable Registration via Attribute Matching and Mutual-Saliency weighting) and the block-matching strategy in [15, 16].

Using intensity-based methods, for example Demons [17], Elastix [18] and Ants [19], is a popular choice in medical applications due to their capability of producing accurate registrations, even between images of different modalities. Unfortunately, intensity-based registration methods tend to be computationally demanding and sensitive to initialization; the objective functions are usually computed over the entire image domain and optimized locally (increasing the risk of getting trapped in a sub-optimal local minimum).

Feature-based methods, using sparse point correspondences between images for establishing coordinate transformations, are typically faster and more robust to initialization and large deformations. The objective functions are typically quantifying residual errors of the mapped point correspondences. This class of objective functions enables efficient computations and optimization methods able to find a global (approximate) minimum. However, these methods risk failing due to the difficulty in detecting salient features in medical images: distinctive features are crucial for establishing correct point-to-point correspondences between the images. Therefore, the accuracy of (sparse) feature-based methods is generally assumed to be inferior to intensity-based methods.

Intensity-based image registration

Intensity-based registration methods rely on comparing pixelvise characteristics such as intensities, colors, depths *etc*. directly. Typically, these methods use local optimization or multiresolution strategies for minimizing an objective function such as sum of squared distances (SSD), sum of absolute distances (SAD), cross-correlation or (normalized) mutual information, (N)MI, [20]. See the comparisons in [21, 22] for different optimization strategies. The non-rigid transformation is commonly represented by deformations derived from physical models, such as the diffusion model in [23] (Demons) or diffeomorphic mapping [24, 25], or by interpolation-based models such as radial basis functions, *e.g*. thin plate splines (TPS) [26], or free-form deformations, *e.g*. cubic B-splines [27]. However, there are numerous nonlinear deformation models in the image registration literature, see the survey in [8].

Feature-based image registration

Despite being a popular choice in computer vision and remote sensing, feature-based image registration is less common in medical image analysis due to the difficulty of detecting distinctive features in medical images. However, Svärm *et al*. [28] showed that feature-based registration based on robust optimization outperforms several intensity-based methods when applied to whole-body CT and brain MRI.

Sparse feature-based registration methods rely on established point-to-point correspondences between images for estimating coordinate transformations. The procedure of establishing point-to-point correspondences includes (i) detection of distinctive feature points in each image and (ii) matching the detected feature points by taking their similarity in appearance into account.

There are numerous hand-crafted feature detectors where the prime examples are Sift [29] (using difference-of-Gaussians) and Surf [30] (using integral images). Detected features are paired with a descriptor, a histogram aiming to provide a

unique description of the feature point and its neighbourhood. These descriptors are computed locally and include image characteristics such as intensity information, gradients, higher order derivatives and/or wavelets. Preferably, the descriptor should be invariant to scale, pose, contrast and, for some applications, rotation. Recently, feature detectors and descriptors learned with convolutional neural networks have proved to excel at several applications [31–33].

Once having detected and described a set of features points for the images that are to be registered, the descriptors need to be matched, in a robust manner, in order to derive correct point-to-point correspondences. Usually, a metric measuring the distance (for example Euclidean distance) between the descriptors is used to rank the quality of match hypotheses. A one-to-one correspondence is derived by choosing the nearest neighbour in the descriptor space (either computed in one direction, non-symmetrically, or compute in both directions, symmetrically), perhaps combined with a criterion such as in [29] (comparing ratios between nearest and second nearest neighbour). Recently, convolutional neural networks have been used for matching as well [34, 35].

Given the correspondence hypotheses, robust optimization algorithms such as Ransac [36] is used to estimate the parameters of a linear transformation approximately, and to sort of out matches that are inconsistent with this linear transformation, outliers. Ransac is typically followed by a global, or a local iterative, optimization procedure using only the inliers, that is, the matches deemed correct by Ransac. A succeeding non-rigid deformation may be represented by interpolation-based techniques, such as B-splines as in [37] or thin plate splines as in [38]. There are also methods simultaneously establishing one-to-one point correspondences while estimating the mapping, such as modified variants of the Iterative Closest Point (ICP) method [39], see the registration method in [40].

2.3 Medical image segmentation

To segment an image means dividing an image into meaningful parts by assigning each pixel to an object class. The classes are a predefined set of objects relevant for the application, such as ”kidney”, ”pancreas”, ”liver” *etc.* for abdominal organ segmentation. The output from a segmentation algorithm is an image labelling, that is, an image of the same dimension as the input image where each pixel has been assigned a label indicating which object class the specific pixel belongs to. A manual labelling, delineated by a physician or other medical expert, is usually referred to as the ground truth labelling. In medical applications, the term gold standard is sometimes used instead (indicating the lack of objective truth when it comes to medical image segmentation).

Image segmentation algorithms aim to find an image labelling, $\mathcal{L}$, that is as similar to the ground truth labelling, $\mathcal{L}_{\text{GT}}$, as possible, that is,

$$\mathcal{L}^* = \arg\max_{\mathcal{L}} S(\mathcal{L}, \mathcal{L}_{\text{GT}}), \tag{2.5}$$

where S is a metric measuring the similarity between two labellings. Segmentation algorithms are typically tuned, or trained, to solve the optimization problem in Equation (2.5) for training images, for which the ground truth labellings are known. Note that the ground truth labellings are unknown for test (evaluation) images. There are several similarity metrics commonly used to train and evaluate segmentation algorithms. One common choice is the Dice coefficient (F1 score), defined as

$$S_{\text{DICE}} = \frac{2|\mathcal{L} \cap \mathcal{L}_{\text{GT}}|}{|\mathcal{L}| + |\mathcal{L}_{\text{GT}}|}, \tag{2.6}$$

where $\mathcal{L}$ and $\mathcal{L}_{\text{GT}}$ are binary labellings for one class. For multi-label problems, the mean Dice metric over all classes is typically used. Another similar metric is the Jaccard index (Intersection over Union), defined as

$$S_{\text{JACCARD}} = \frac{|\mathcal{L} \cap \mathcal{L}_{\text{GT}}|}{|\mathcal{L} \cup \mathcal{L}_{\text{GT}}|}. \tag{2.7}$$

The relation between the two metrics is $S_{\text{DICE}} = 2S_{\text{JACCARD}}/(1+S_{\text{JACCARD}})$. Both metrics have values between zero and one, where higher means better. There are several other similarity metrics in the literature, where the Hausdorff distance and the mean surface distance are two examples used in applications where the qualitative segmentation shape is important.

There are numerous different segmentation algorithms based on thresholding, region growing, edge detection, variational methods, level sets or shape models. In this section, two commonly used methods for medical applications, using implicit shape modelling, are summarized.

2.3.1 Multi-atlas segmentation

Multi-atlas segmentation [41–43], proposed over a decade ago, is one of the most widely used methods for segmentation in medical applications. For an extensive summary of the research field, see the survey in [10].

Multi-atlas segmentation is an extension of single-atlas segmentation. An atlas is an image paired with a corresponding ground truth labelling. Single-atlas segmentation relies on registering one atlas image to the unlabelled target image and transferring the labelling according to the computed transformation. Thus, the inferred target image segmentation equals the aligned labelling. For that reason, single-atlas segmentation is also called registration-based segmentation, see Figure 2.1.

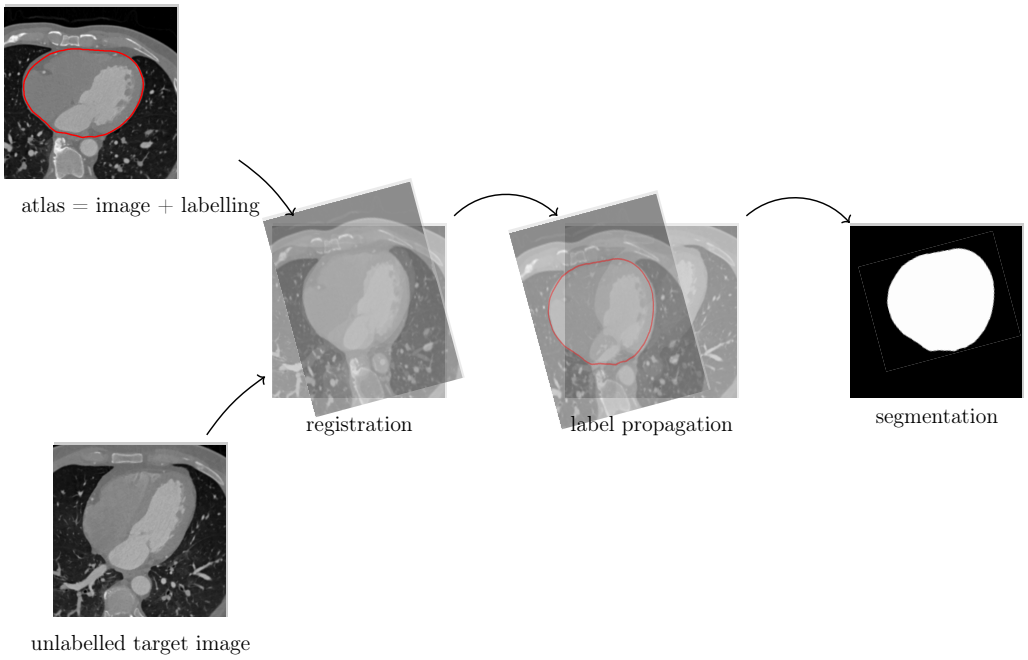
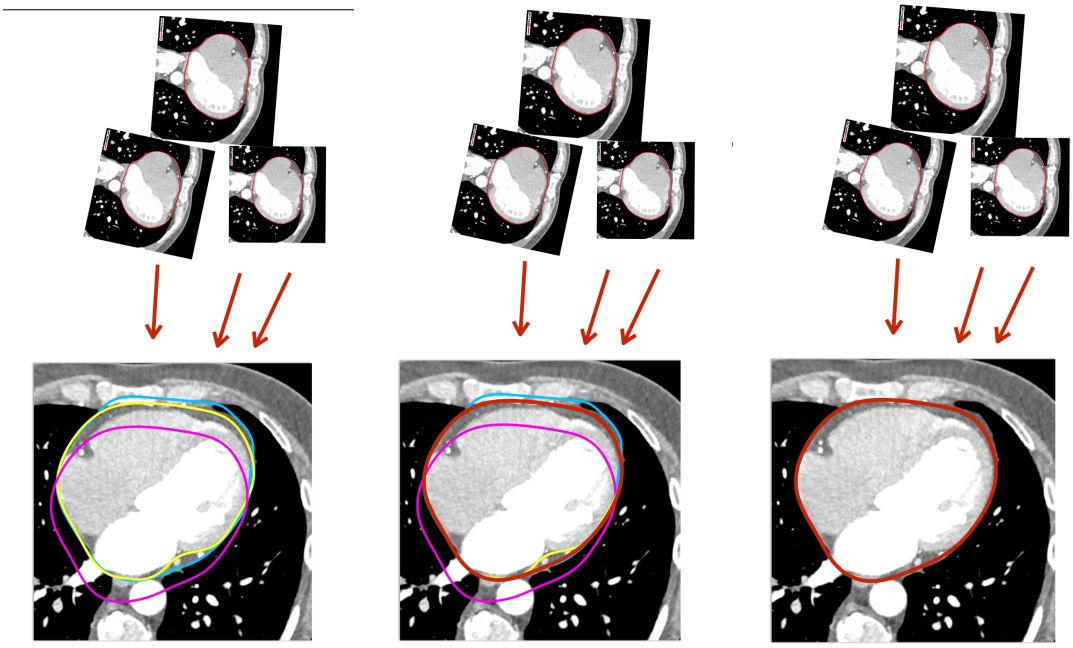


Figure 2.1: *Example of single-atlas segmentation (registration-based segmentation) of the pericardium in a* Scapis *cardiac CTA slice.*

Two or more single-atlas segmentations can be combined into a multi-atlas segmentation. The motivation behind using several atlases is to capture more possible anatomical variations and to increase the robustness to imperfect registration results. Thus, multi-atlas segmentation involves registration of several atlas images to the unlabelled target image. According to the pairwise atlas-target registrations, each atlas labelling is propagated to the target image space and thereafter combined via label fusion, see below. Figure 2.2 depicts an example of a coarse multi-atlas segmentation (Scapis pericardium segmentation) using three atlases.

Label fusion

In multi-atlas segmentation, there are several propagated atlas labellings that need to be combined into one unique segmentation proposal. Each transferred atlas labelling can viewed as a vote, for each pixel indicating whether that particular atlas estimates the pixel to be inside/at the organ boundary or not. By summarizing all votes in one image a voting map is obtained. The voting map can be regarded as an unnormalized pixelwise label likelihood over the entire image. From this voting map, the final segmentation can be inferred by, for instance, thresholding or statistical reasoning. The process of combining several transferred atlas labellings into one voting map is referred to as label fusion. For some label fusion schemes, the



(a) (b) (c)

Figure 2.2: *Example of a multi-atlas segmentation of the pericardium in a slice of a* Scapis *cardiac CTA image using three atlases. (a) The atlas images are registered to the unlabelled image and the labellings (coloured contours) are transferred accordingly. (b) The transferred labellings are combined into one segmentation proposal (red contour) by label fusion. (c) The inferred segmentation accurately delineates the pericardium compared to the individual single-atlas segmentations.*

output simply equals the voting map, that may be used in a subsequent analysis step, while other fusion strategies output the final inferred segmentation proposal.

The simplest fusion scheme is unweighted voting [41–43], meaning that each registered atlas is assigned the same weight, see Figure 2.3c. Typically, methods using unweighted voting maps infer the final segmentation by majority voting, that is, the most frequent label is assigned to each pixel.

It is common to sift out promising atlas candidates and only fuse this restricted subset. This process, known as atlas selection, has proven to improve the computational efficiency (by decreasing the amount of registrations that need to be computed) and accuracy (by ignoring irrelevant anatomies), see Figure 2.3e. Atlas selection can be done either before pairwise registration, as in [44], by choosing atlas images believed to best represent the anatomical shape variation, or after, as in [45], by choosing the atlas images which are more similar to the target image and/or are believed to boost the algorithm performance. The most simple case of atlas selection is best atlas selection [41], where merely one atlas is chosen, see

Figure 2.3f. Atlas selection may be regarded as an extreme case of weighted voting, that is, fusing propagated labels by assigning each atlas different weights, see Figure 2.3d. The atlas weights can be derived globally, as in [45, 46], or locally (patchwise or pixelwise) as in [47–51].

There are numerous additional sophisticated fusion schemes including ideas from statistics and machine learning. Among others, there are strategies using probabilistic reasoning regarding predicted performance [45, 52–54], generative probabilistic models [55] and convolutional neural networks [56].

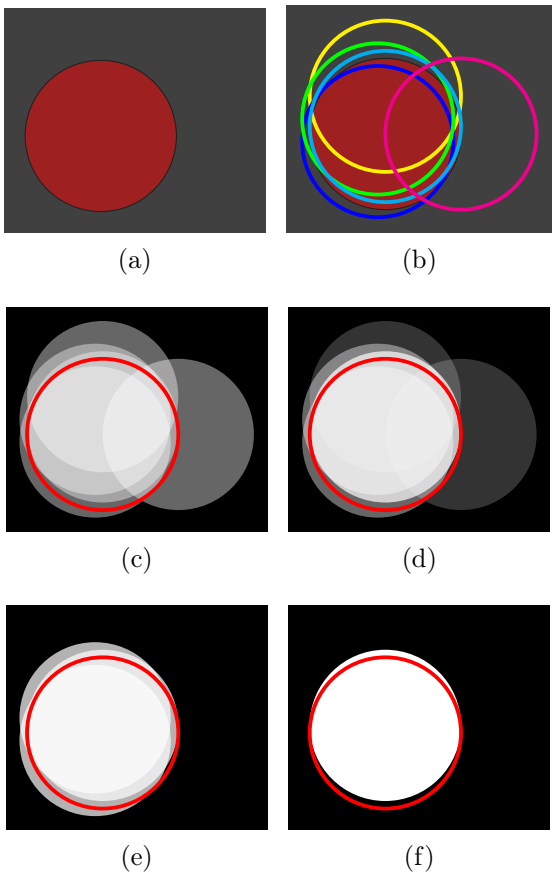


Figure 2.3: *Toy example visualizing different label fusion strategies. (a) An unlabelled image depicting a red, circular shape on a gray background. (b) Five atlases are registered to the unlabelled image and labellings (coloured contours) are propagated accordingly. (c) Unweighted voting assigns the exact same weight to each atlas. The red contour represents the true boundary. (d) Weighted voting assigns different weights to each atlas. (e) Atlas selection sifts out promising atlas candidates. (f) Best atlas selection sifts out the most promising atlas candidate.*

2.3.2 Conditional random fields

Conditional random fields (CRFs), a variant of Markov random fields (MRFs) [57–59], is a class of probabilistic graphical models suitable for modeling spatial context such as smooth segmentation boundaries, coherent shapes *etc*. CRFs may be regarded as implicit shape models; they do not directly enforce an explicit (parameterized) shape model but still encourage spatial smoothness between neighbouring pixels. By also considering the classification of neighbours when assigning a label to a pixel, noisy or implausible boundaries can be avoided. CRFs have successfully been used for medical image segmentation [60–63], see the survey in [64].

When using CRFs for computing segmentations, the labelling problem is posed as an optimization problem that is solved either exactly (if possible) or approximately. More specifically, the image is regarded as an observation of a conditional random field and the labelling (the realization of the field) is inferred by solving an energy minimization problem.

Mathematical model

Let $l_p \in \mathcal{L}$ be a variable indicating what class a pixel, indexed by $p \in \mathcal{P}$, is assigned to and let $i_p \in \mathcal{I}$ denote the observed intensity for the pixel. Here, $\mathcal{I}$ denotes the image, $\mathcal{L}$ denotes the labelling and $\mathcal{P}$ denotes the set of all pixel indices. The optimal segmentation is inferred as the labelling that maximizes the posterior probability given by

$$P(\mathcal{L} \mid \mathcal{I}; \boldsymbol{\theta}) = \frac{1}{Z} e^{-E(\mathcal{L}, \mathcal{I}; \boldsymbol{\theta})}, \tag{2.8}$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \ldots)$ are tunable parameters and Z is the partition function (the normalizing constant). The parameters are either fixed (*e.g.* derived by prior assumptions) or learned during training.

In most image applications, the energy E is assumed to decompose over unary and pairwise potentials. If so, the energy can be expressed as

$$E(\mathcal{L}, \mathcal{I}; \boldsymbol{\theta}) = \sum_{p \in \mathcal{P}} \phi_p(l_p, \mathcal{I}; \boldsymbol{\theta}) + \sum_{(p,q) \in \mathcal{N}} \phi_{p,q}(l_p, l_q, \mathcal{I}; \boldsymbol{\theta}), \tag{2.9}$$

where the set of all pairwise neighbours is denoted as $\mathcal{N}$. The unary potential ϕ_p may also be referred to as the unary cost, unary energy or data cost. Similarly, the pairwise potential $\phi_{p,q}$ may be referred to as the pairwise cost, pairwise energy or regularization/coherence cost. In some applications, it may be beneficial to include potentials of higher orders (cliques including three or more neighbours), as in [65].

The neighbourhood of a pixel is defined by the pixel connectivity. In 2D applications, common choices are 4-connectivity (neighbours are defined by connected

edges) and 8-connectivity (neighbours are defined by connected edges and corners). For 3D, common choices are 6-connectivity (neighbours are defined by connected faces), 18-connectivity (neighbours are defined by connected faces and edges) or 26-connectivity (neighbours are defined by connected faces, edges and corners). However, larger neighbourhoods are also allowed. Further, one may incorporate the distance between pixels directly in the potentials, letting the pairwise energy depend smoothly on pixel distances (dense CRFs). If so, the second term in Equation (2.9) is summarized over all possible pixel combinations.

The unary cost, also known as the data cost, is usually dependent on conditional probabilities learned from data, such as the label likelihoods computed by a multi-atlas voting map or a machine learning classifier. A typical choice is

$$\phi_p = \theta_1 \log(\hat{P}(l_p \mid \mathcal{I})), \tag{2.10}$$

where $\hat{P}(l_p \mid \mathcal{I})$ equals the previously estimated likelihood.

The pairwise cost is an interaction term that regularizes the solution. In the simplest case, the pairwise costs are set to a fixed constant for all neighbours assigned with different labels, neighbours with the same labels are not penalized. This is called a Potts model:

$$\phi_{p,q} = \mathbb{1}_{l_p \neq l_q} \theta_2, \tag{2.11}$$

where $\mathbb{1}_{l_p \neq l_q}$ denotes the indicator function equaling one if $l_p \neq l_q$, that is, if the neighbours are assigned different labels. However, more complex pairwise potentials taking the neighbouring intensities into account as well are usually beneficial. A common choice of the pairwise energy, consisting of two terms both penalizing neighbouring pixels being labelled differently, is given by

$$\phi_{p,q} = \mathbb{1}_{l_p \neq l_q} \left(\theta_2 + \theta_3 e^{-d(i_p, i_q)} \right), \tag{2.12}$$

where $d(\cdot, \cdot)$ is a metric measuring *e.g.* the contrast of the neighbouring pixels.

Unfortunately, the pairwise interaction term may lead to a bias towards shorter segmentation boundaries, a shrinking bias. However, there are several proposed solutions in the literature, *cf.* [66, 67]

Inference

A function on the form in Equation (2.9) can be formulated as a weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of nodes (pixels) and $\mathcal{E}$ is the set of edges connecting neighbouring pixels. If the segmentation problem is binary and if the energy in Equation (2.9) is submodular, the globally optimal labelling can be computed exactly and in polynomial time using graph cuts [68]. Otherwise, methods such as alpha expansion [69], mean field inference or linear programming relaxations may be used to solve the minimization problem approximately.

2.4 Machine learning for medical images

The last couple of decades, the field of machine learning has provided algorithms excelling at computer vision tasks. Along this development, machine learning tools for image classification and regression have received a great deal of attention from the medical image analysis community. Hand-crafted features and models have successfully been replaced with learned equivalents in segmentation and registration tasks. The increased interest and prosperity can predominantly be explained by improved computer hardware and the increased access to large annotated medical image datasets [70].

Included thesis papers make use of two types of machine learning tools, random decision forests and convolutional neural networks. Therefore, a brief overview of the techniques follows below.

2.4.1 Random decision forests

Random decision forests [71, 72] (short: random forests) are a machine learning technique suitable for classification and regression tasks. It is a computationally efficient method and it generalizes well to unseen data. In the field of medical image analysis, random forests have been applied to both registration tasks, *e.g.* abdominal CT [73–75], spine CT [73, 75], whole-body CT [73] and brain MRI [76, 77], as well as segmentation tasks, *e.g.* pelvic radiographs [78], cardiac and abdominal MRI [79], brain MRI [80–82], pelvic CT [83–85], abdominal CT [83, 84, 86–88], cardiac and pulmonary CT [83, 89, 90] and femur ultrasound [81].

For segmentation tasks, random decision forests typically estimate pixelwise probabilities for each label, that is, a likelihood estimate for each pixel belonging to a certain class. When applied to an unlabelled pixel, the random decision forest is fed a set of features, *i.e.* characteristics derived from the image, as input and outputs an estimated conditional probability over labels, $\hat{P}(l|\mathbf{f})$, where l denotes the pixel label and $\mathbf{f}$ denotes a vector consisting of the input features. The output labelling may be found by maximizing the output distribution, or by feeding the posterior distribution as a data term to a conditional random field model, see Section 2.3.2.

Some of the listed applications use regression forests, instead of the classification forests described above. The principles for regression forests are similar, however, the prediction is instead computed as the mean of the output posterior distribution. The included thesis papers use classification forests exclusively. Therefore, classification forests are used as the running example in the detailed description below.

Decision trees

A random decision forest consists of a set of decision trees, binary trees where each node is associated with its own splitting (decision) function. A common choice of splitting function is a separating hyperplane of the same dimension as the input feature vector. The parameters of the hyperplane are learned during training and usually chosen such that the information gain (the confidence) is maximized and/or the entropy (the unpredictability) is minimized.

The purpose of the splitting function is to separate the input data points based on feature similarity. Typically, features such as image intensities, gradients and/or higher order derivatives are used. It is also common to pre-process the image, for example by filtering, and include these pre-processed intensities as features. It is good practice to normalize each feature before training to have zero mean and unit standard deviation with respect to the training set.

When classifying an unlabelled pixel, the input data point begins at the rote node. Depending on the result of the current splitting function (the decision), the data point is either passed to the right or to the left child node. The subsequent nodes will continue passing the data point along the tree until it reaches a leaf node. The leaf nodes contain posterior distributions over labels, learned during training, and thus output a conditional probability for the data point belonging to a certain class.

In Figure 2.4a training of a binary decision tree is visualized. In this specific example, 20 data points are used for training. There are two classes, blue and red, and two different features have been extracted for each data point. That is, the classification problem is two-dimensional. The binary decision tree has in total six nodes: one root node, two decision nodes and three leaf nodes. Below the leaf nodes, the estimated posterior distribution for the two different classes (for that particular leaf) is given.

In Figure 2.4b classification of one unlabelled data point is visualized. The data point is passed along the three according to the decision nodes, and the estimated posterior distribution over the classes is decided by the leaf node the data point end up in. For this particular example, the data point would be classified as ”red”, since the estimated posterior distribution is the largest for this class.

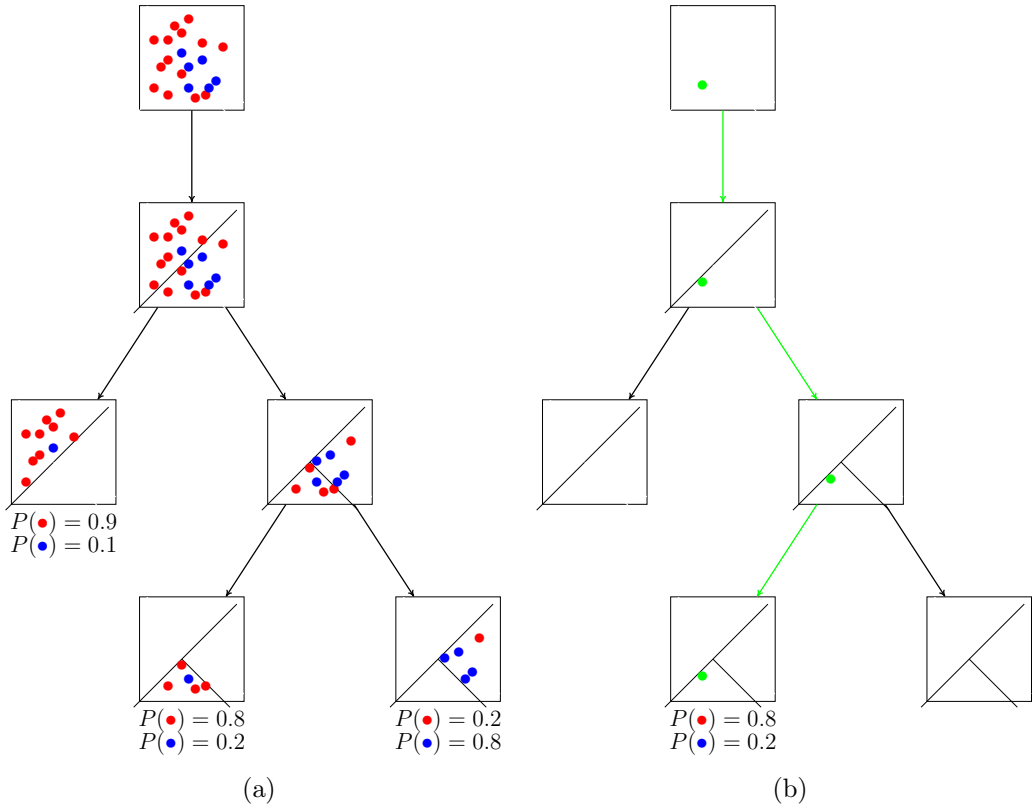


Figure 2.4: *Example of a binary decision tree consisting of six nodes; one root node, two decision nodes and three leaf nodes. (a) The decision tree is trained on 20 data points belonging to two different classes, "red" and "blue". For each data point, two different features have been computed. The two decision nodes (containing splitting functions equaling separating hyperplanes) are trained to divide the data into three different distributions (the leaf nodes). Each leaf node provides a posterior distribution over the classes for test data points ending up in that particular leaf node. (b) Features for an unlabelled data point (green) are computed and the data point is passed along the decision tree according to the splitting functions. The unlabelled data point ends up in the middle leaf node and is thus classified as "red".*

Random forests

Decision trees tend to overfit training data, that is, they have a low bias but a high variance. Therefore, random forests consist of several decision trees where each decision tree is trained on a random subset of the training data (referred to as tree bagging). The estimated posterior probability is typically computed as the average over all trees:

$$\hat{P}(l|\mathbf{f}) = \frac{1}{T}\sum_{t=1}^{T}\hat{P}_t(l|\mathbf{f}), \tag{2.13}$$

where l denotes the label, $\mathbf{f}$ denotes the feature vector and T equals the number of trees. To further reduce variance by decorrelating the trees, only a subset of the features is randomly chosen at each tree node.

2.4.2 Convolutional neural networks

Convolutional neural networks (CNNs) constitute a class of machine learning tools for classification and regression in image, video and natural language processing. Despite being introduced already in the 70s [91] by the name "Neocognitron", CNNs have received a great deal of attention from the image analysis and computer vision research community the last decade. The popularity stems from recent success on problems such as image classification [92] and object detection [93]. The success can predominantly be explained by an increased computational power of modern GPUs (Graphical Processing Units) and the access to large annotated datasets. Below follows a brief introduction to the technique, see the overview in [94] for more details.

Due to their outstanding results on a wide variety of tasks and applications, CNN-based methods have emerged in the field of medical image analysis as well. So far, CNNs have been applied to segmentation of *e.g.* electron microscopy images [95, 96], knee MRI [97], prostate MRI [98], abdominal CT [99, 100], spine MRI [101], cardiac MRI [102] and brain MRI [103–106], as well as registration of *e.g.* brain MRI [107–111], pulmonary CT [112, 113], cardiac MRI [114–116] and multi-modal MRI/ultrasound [117].

CNNs are feed-forward artificial networks consisting of trailing computational layers where connections enable the result from one layer to be forwarded to a subsequent layer for further processing, see Figure 2.5. CNNs are universal function approximators, that is, the they are (in theory) able to model any function. To enable this capacity, the computational layers contain thousands or millions of parameters that are automatically learned during training.

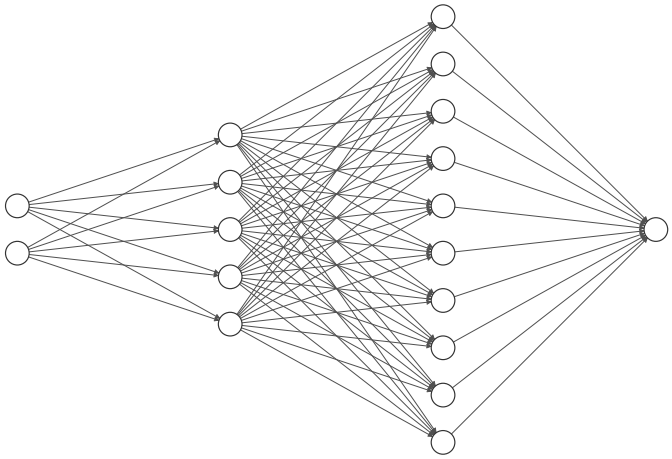


Figure 2.5: *An example of a feed-forward artificial network with an input layer consisting of two input units, two hidden layers consisting of five and ten computational units respectively, and an output layer consisting of one output unit.*

Computational layers

A simple CNN consists of one input layer, one output layer and one or more hidden layers. The input layer usually equals a full image, however, other input layers such as smaller input patches are also common depending on the network architecture. In contrast to other image analysis algorithms, pre-processing of the input data is typically not required when using CNNs since any needed image processing is learned automatically. In CNNs constructed for classification or segmentation problems, the output layer typically equals conditional probabilities over predefined object classes, *cf*. the output of random decision forests in Section 2.4.1. For image classification problems, the CNN outputs a likelihood for image subjects, for example whether the image depicts a dog, a cat or a horse. CNN constructed for pixelwise classification, such as segmentation networks, instead outputs label likelihoods for each pixel. For image regression problems, the CNN outputs image- or pixelwise predictions, depending on the task at hand.

The purpose of the hidden layers is to map the given input to the desired output. To enable modeling of any complex function, the hidden layers contain several different building blocks such as sets of learnable filters (convolutional layers), downsampling layers (pooling layers) and decision functions (nonlinear activation functions). Typically, CNNs consist of a set of trailing convolutional layers terminated with nonlinearities and layered with pooling layers. However, there are numerous proposed architectures in the literature. It is generally assumed that networks containing many small convolutional layers (deep networks) are more likely to produce good results than networks containing a few large convolutional layers (wide, shallow networks), but the findings so far are inconclusive [118].

Convolutional layers. The purpose of the convolutional layers is to extract image characteristics with means of automatically learned filters. Each convolutional layer typically contains several learnable filters, filter banks. The output from each filter, called the filter response or the feature map, is forwarded to succeeding layers for further processing. Ideally, the first few convolutional layers extract low-level features, such as blobs, edges, corners, lines *etc*., while later layers combine these low-level features into more complex features such as human faces. The depth and the width of the network, that is, the amount of subsequent layers and their size, decide the learned filters ability to recognize high-level features. In contrast to hand-crafted feature detectors and descriptors such as Sift or Surf, the CNN filter parameters (filter weights) are automatically learned during training and thus not designed with any prior knowledge in mind. The convolutional property enables translation invariance, that is, input patterns in different parts of the image is processed in the exact same manner. Dilated convolutions [119] and non-unit filter strides are two common strategies to increase the receptive field, *i.e*. the region of the input image that is visible to each filter.

Pooling layers. The pooling layers aim to downsample the image (and subsequent filter responses) in order to reduce the parameter space preventing undesired effects such as overfitting and unnecessary high computational complexity. By downsampling, the pooling layers also introduce non-linearity. Two common choices of pooling is max pooling, by applying a maximum filter, and average pooling, by applying a mean filter. Note that pooling layers in principle equal convolutional layers with fixed (non-learnable) filter weights. As for convolutional layers, dilated pooling and non-unit filters stride may help increasing the receptive field.

Non-linear activation functions. Non-linearities are important to enable the universal function approximator property; using only linear combinations of convolutional layers would enable nothing but linear maps from input to output. The non-linearities also restrict unbounded layer outputs to a certain range, and thus help avoiding an accumulation of large values in some sections of the network. There is a wide selection of activation functions such as the rectified linear units (ReLUs) [120], sigmoid units and tangens hyperbolicus units. In modern networks, ReLU or its variants (leaky ReLU [121], parametric ReLU [122] and Swish [123]) are the most popular choices. The nonlinear softmax unit, mapping real numbers to probabilities, is particularly useful in the output layer of classification/segmentation networks.

Fully connected layers. Before the output layer, there are sometimes one, two or more fully connected layers. The fully connected layers aim to map a large set of multidimensional filter responses to a more manageable 1D histogram. For instance, a CNN constructed for distinguishing two image classes typically terminates with fully connected layers mapping the filter responses to a histogram of size two. Applying the softmax operator to this histogram gives a conditional probability estimate for the two classes.

Fully convolutional networks

CNNs including fully connected layers are not particularly efficient when dealing with pixelvise classification (or regression) tasks; these networks can not be trained on nor be applied to images of arbitrary sizes. Moreover, the fully connected layers have a large amount of parameters and are computationally demanding.

Another class of networks, fully convolutional networks [96, 98, 104, 124, 125], is better suited for tasks requiring pixelwise outputs. Fully convolutional networks drop the terminating fully connected layers. Instead, they solely use convolutional and pooling layers for filtering and downsampling the image. These networks are capable of processing images of arbitrary sizes, and they are computationally more efficient than their fully connected counterparts. To enable outputs of the same size as the input, some fully convolutional networks include deconvolution layers. The purpose of the deconvolution layers is to upsample and merge the filter responses from earlier layers, enabling dense pixelwise predictions. Fully convolutional networks having this structure of filtering/downsampling and "de-filtering"/upsampling the image are called encoder-decoder networks. To avoid loosing spatial information due to pooling, these networks typically process the features at different resolutions, and/or replace the pooling layers entirely with dilated convolutions and non-unit filter strides. See Figure 2.6 for an example of an encoder-decoder network.

Learning

The convolutional layers of a CNN consists of a huge amount of parameters that need to be learned. Learning is achieved by optimizing an objective function that quantifies the compatibility of the network's output and the desired output (such as the ground truth labelling for segmentation tasks).

CNNs are trained using local optimization methods, common choices are stochastic gradient descent or mini-batch gradient descent. To speed up convergence, there are variants using batch normalization [126], Nesterov's momentum [127] and adaptive learning rate (*e.g.* AdaGrad [128], RMSprop/AdaDelta [129], Adam [130], Nadam [131]). Despite complex architectures and a huge amount of parameters, the gradients can be efficiently computed using the backpropagation algorithm,

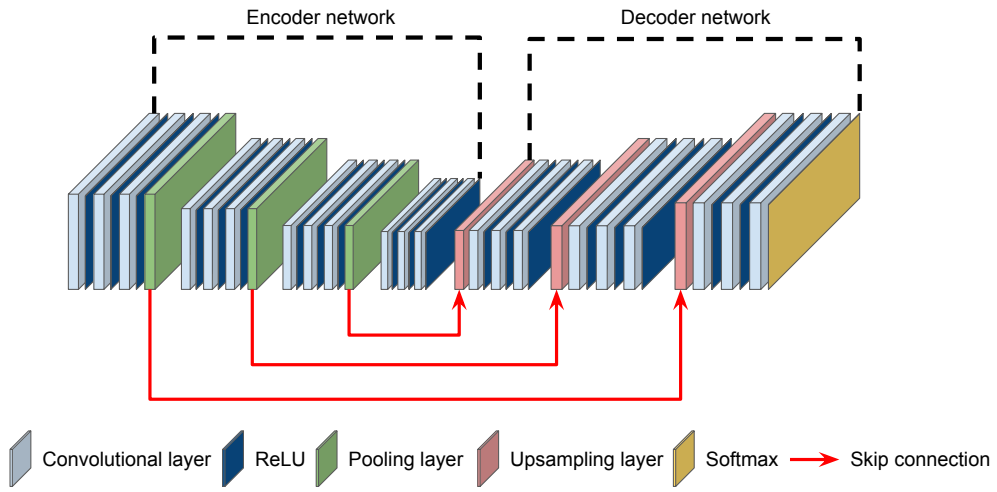


Figure 2.6: *An example of a fully convolutional encoder-decoder network consisting of convolutional layers, ReLU activations, pooling layers, upsampling layers and a terminating softmax layer. The skip connections enable forwarding of features from early to late layers, in order to avoid loosing information necessary for the reconstruction in the down-sampling phase.*

first proposed in [132–134]. Training is done in epochs, where all training samples are utilized in each epoch. For classification networks using a terminating softmax unit, pixelwise cross-entropy is commonly used as objective function. Another choice of objective function is the max-margin hinge loss allowing for a support vector machine (SVM) classifier. For regression networks, mean square error and mean absolute error are the most common choices of objective function.

Due to the large amount of learnable parameters, an important consideration during network training is to prevent overfitting. An overfitting network performs well on training data, but fails to generalize to unseen data. There are several techniques for this, such as batch normalization (mentioned above), dropout [135], filter weight regularization and early stopping [136]. Ideally, overfitting is solved by presenting a sufficient amount of training examples to the network. However, manually labelled data is rarely abundant. The training data can be artificially augmented by adding small random perturbations to the training samples, such as rotations, additive noise, scaling *etc*. When faced with a new task, it can be beneficial to use a pre-trained CNN, especially if training data is limited. Pre-training can be done either using other (preferably similar) datasets or with means of unsupervised training as in [137]. Pre-training facilitates learning by enabling the network to re-use filters that have already learned to recognize certain low-level features.

Chapter 3

Thesis contributions

As detailed in Section 1.1, excellent medical registration and segmentation algorithms are characterized by speed, allowing for scalability, and an accuracy comparable to an expert radiologist. They should allow for plausible organ (or region) shapes while generalizing well to unseen and rarely occurring anatomies. Preferably, training the algorithms should be data-efficient since manually labelled data typically is scarce in the medical community. Thus, these are all aspects considered in the included papers:

Paper I mainly concerns increasing image registration speed, accuracy and generalizability with means of a CNN.

Paper II mainly concerns improving the multi-atlas segmentation pipeline taking plausible organ shapes into account.

Paper III mainly concerns improving CNN segmentation results by incorporating a CRF model ensuring plausible boundaries.

Paper IV mainly concerns improving a multi-atlas segmentation framework paired with a random decision forest classifier with respect to accuracy and data-efficiency.

Paper V mainly concerns speeding up a feature-based image registration procedure via clustering and robust optimization.

This chapter is structured as follows: each section constitutes an overview of one of the included thesis papers including a summary of the main algorithmic contributions. Also, the contributions of the thesis author are stated for each paper respectively.

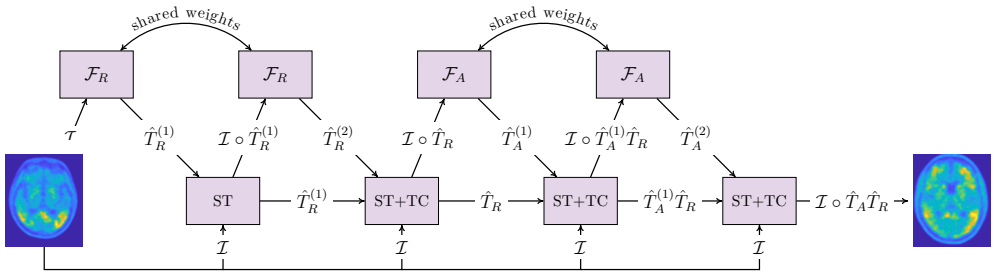


Figure 3.1: *Schematic illustration of the implemented network in paper I, see paper for more details.*

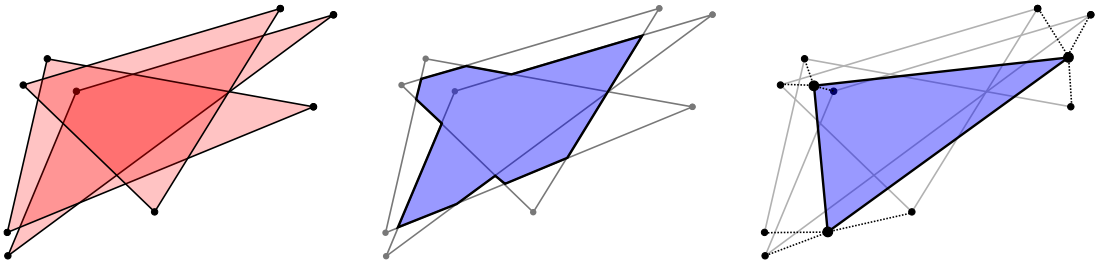
3.1 Paper I

J. Alvén, K. Heurling, R. Smith, O. Strandberg, M. Schöll, O. Hansson and F. Kahl. "A Deep Learning Approach to MR-less Spatial Normalization for Tau PET Images". *The International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2019.

The procedure of aligning a subject's PET image with a common MR template is called spatial normalization, and is essential for PET analysis. Most approaches to spatial normalization align the PET image with the template space via a MR image of the same subject. One major disadvantage is the need for the subject's MR, and enabling PET spatial normalization without MR would most definitely benefit large-scale studies. Common for all previous attempts on spatial normalization without MR is the use of standard image registration techniques with an explicit PET template as target. However, such template models do not always capture the full variation of PETs, which makes these methods less robust and unreliable for general PET images.

This paper proposes a method that aligns the PET image directly without MR, and without using an explicit PET template. A deep neural network estimates an aligning transformation from the PET input image, and outputs the spatially normalized image as well as the parameterized transformation. In order to do so, the proposed network iteratively estimates a set of rigid and affine transformations by means of convolutional neural network regressors as well as spatial transformer layers, and is trainable end-to-end.

Author contribution. I implemented the full method, run all experiments and wrote large parts of the paper. Kahl and Schöll helped with the writing. I, Kahl, Heurling and Schöll proposed the main idea. Smith, Strandberg and Hansson acquired the BioFinder data.



(a) Triangles (b) Majority voting (c) Shape averaging

Figure 3.2: *The concept behind the shape averaging in paper II.*

3.2 Paper II

J. Alvén, F. Kahl, M. Landgren, V. Larsson, J. Ulén and O. Enqvist. "Shape-Aware Label Fusion for Multi-Atlas Frameworks". *Pattern Recognition Letters,* 124:109-117, 2019.

Good segmentation algorithms should generalize well to unseen or rarely occurring anatomies while still producing plausible organ (or region) shapes. Multi-atlas segmentation frameworks tend to generalize well, also when labelled training data is limited. However, standard multi-atlas segmentation methods puts no explicit constraints on the output shape. On the contrary, standard multi-atlas label fusion combines transferred labels locally by merely considering the current voxel and/or spatially neighbouring voxels. In order to guarantee a preserved topology and to prevent disjoint organ shapes or lost structures, global shape regularization needs to be included. Unfortunately, most methods with explicit shape constraints fail generalizing as well as multi-atlas methods do.

This paper incorporates a shape prior into multi-atlas label fusion without losing the generalizability of multi-atlas methods. Instead of fusing the labels at the voxel level, each transferred labelling is regarded as a shape model estimate. The shape model is a point distribution model of the organ surface consisting of landmark correspondences established offline. Online, pairwise registrations provide coordinate estimates for these landmarks in the target image. These estimates are used for computing an average shape by using robust optimization techniques. In this manner, an awareness of the overall shape is directly incorporated into the label fusion preventing implausible results while keeping robustness to outlier registrations. See Figure 3.2 for a visualization of the concept of shape averaging.

Author contribution. Implementations, experiments as well as the writing were joint work. I mostly contributed to (i) implementations related to the CNNs and the landmarks establishment, (ii) running the experiments and (iii) writing the paper. I, Kahl and Enqvist proposed the main idea.

Figure 3.3: *Qualitative results on a* Scapis *CTA sagittal slice from paper III. The red number in the upper right corner is the Jaccard similarity index (%). See paper for more details.*

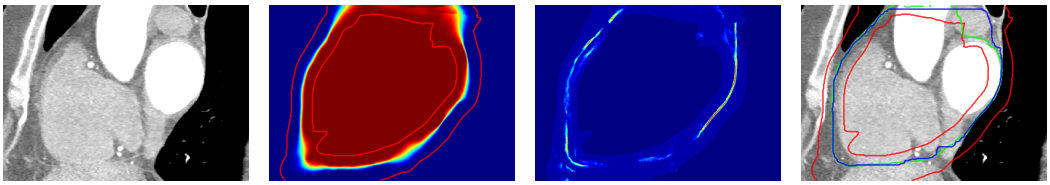
3.3 Paper III

M. Larsson, J. Alvén, and F. Kahl. "Max-margin learning of deep structured models for semantic segmentation." *Scandinavian Conference on Image Analysis (SCIA),* 2017.

Convolutional neural networks have proven powerful for image segmentation tasks, due to their ability to model complex connections between input and output data. However, CNNs lack the ability to model statistical dependencies between output variables, for instance, enforcing properties such as smooth and coherent segmentation boundaries. In order to guarantee plausible segmentation shapes, condition random fields can be used as a post-processing step. However, using CRFs only as a refinement step means that the paired CNN and CRF are trained separately, that is, the parameters of the CRF are learned while the parameters of the CNN are fixed, and vice versa. A better solution is end-to-end learning, where the CNN and CRF parameters are learned jointly.

This paper proposes a learning framework that jointly trains the parameters of a CNN paired with a CRF. In order to do so, a theoretical framework for optimization of a max-margin objective with back-propagation is developed. The max-margin objective ensures good generalization capabilities, which makes the method especially suitable for applications where labelled data is limited, such as medical applications. The method is successfully evaluated on two medical segmentation tasks, pericardium segmentation in Scapis CTA slices and heart ventricle segmentation in Echo ultrasound slices. Figure 3.3 shows a comparison of the piecewise and jointly trained models for a Scapis CTA sagittal slice.

Author contribution. I implemented methods for producing the manual labellings of the Scapis CTA slices and the Echo ultrasound slices. Larsson carried out the algorithm implementations and the experiments. The writing of the paper were joint work, and Kahl proposed the main idea.



(a) (b) (c) (d)

Figure 3.4: *Visualization of the main parts of the algorithm in paper IV. (a) Slice of target volume, (b) multi-atlas distance map, (c) random forest posterior probability and (d) the ground truth (green) and the final segmentation (blue).*

3.4 Paper IV

A. Norlén, J. Alvén, D. Molnar, O. Enqvist, R. Rossi Norrlund, J. Brandberg, G. Bergström and F. Kahl. "Automatic Pericardium Segmentation and Quantification of Epicardial Fat from Computed Tomography Angiography". *Journal of Medical Imaging,* 3(3), 2016.

Voxelwise classification with means of machine learning techniques can improve segmentation results, especially for challenging tasks such as pericardium segmentation. However, machine learning tools are dependent on large sets of labelled data, which are rarely occurring in medical applications. The paper addresses the problem of overcoming a shortage of labelled data when applying a random forest classifier to pericardium segmentation.

The primary algorithmic contribution of this paper is the incorporation of a generalized formulation of multi-atlas segmentation based on distance maps into a random forest classification framework. More specific, transferred atlas labellings define a voxelwise distribution over distances to the organ boundary. This distribution serves as a global initialization for the organ boundary search space. Further, it provides a local coordinate system enabling alignment of extracted features to the organ boundary. Rotation invariant features greatly simplify the voxel classification task (reducing the 3D boundary detection problem to 1D line search) but also normalize the training data leading to more efficient use of the labelled data set. In this manner, the random decision forest classifier learns recognizing organ boundaries irrespective of the orientation relative the image coordinate axes. For a visualization of the main steps, see Figure 3.4.

Author contribution. I carried out all baseline experiments and contributed with some ideas. Norlén carried out most of the algorithm implementations. The rest of the implementations and experiments as well as the writing were joint work. Norlén, Enqvist and Kahl proposed the main idea.

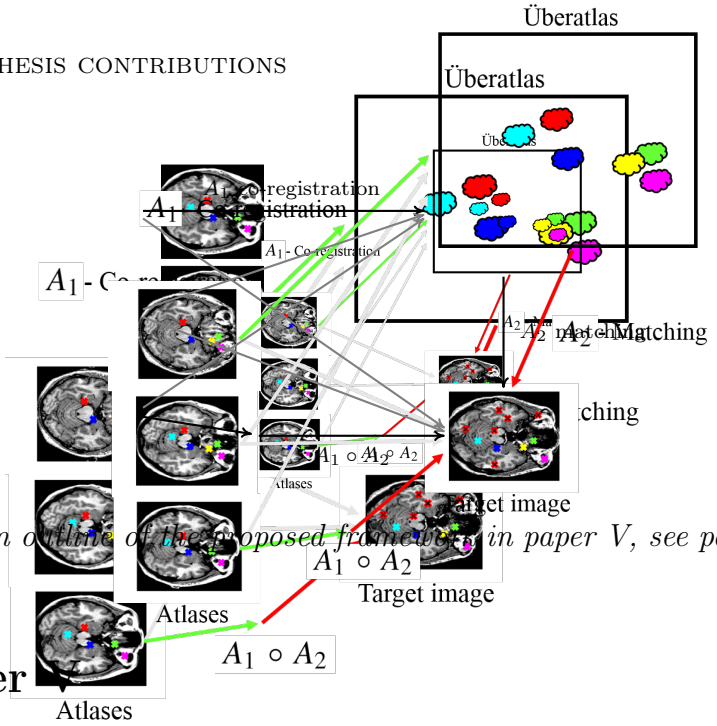


Figure 3.5: *An outline of the proposed framework in paper V, see paper for more details.*

3.5 Paper V

J. Alvén, A. Norlén, O. Enqvist and F. Kahl. "Überatlas: Fast and Robust Registration for Multi-Atlas Segmentation". *Pattern Recognition Letters,* 80:245–255, 2016.

Multi-atlas segmentation has the disadvantage of requiring multiple atlas registrations to capture the full range of possible anatomical variation. In general, image registration is computationally heavy which consequently limits the practical size of the atlas set. To speed up the registration procedure, and thus allowing for larger atlas sets, the paper proposes an intermediate representation of the atlas set. The intermediate representation consists of feature points that are similar and consistently detected throughout the atlas set. This intermediate representation may be used for simultaneously finding point correspondences and affine transformations to a target image from an arbitrarily large set of atlas images.

The main idea is to cluster extracted feature points from the atlas set to form the intermediate representation. To make sure the feature points in a cluster describe the same anatomical feature, the clustering procedure takes both descriptor distances and spatial distances (according to an offline spatial co-registration of the atlases) into account. At running time, point correspondences to all atlas images are automatically obtained at once, and affine transformations can be computed quickly and robustly with means of the iteratively reweighted least squares algorithm. For a schematic overview, see Figure 3.5.

Author contribution. I implemented most of the framework including the clustering algorithm, the atlas co-registration and the iteratively reweighted least squares algorithm. I also run all the Elastix experiments. The remainder of the implementations, experiments as well as the writing were joint work. All authors contributed to the main idea.

Chapter 4

Concluding discussion

The thesis papers propose possible improvements to medical image segmentation and registration algorithms based on shape modelling and machine learning. Improvements are made with respect to accuracy, anatomical/physiological plausibility, speed and scalability as well as generalizability. Below follows a brief discussion regarding how successfully these objectives are handled as well as an introduction to two future research projects each addressing one or several of these objectives.

4.1 Discussion

4.1.1 Accuracy and fair evaluation

All included papers seemingly manage to meet the objective to increase the segmentation and registration accuracy, each paper proposes algorithms performing better or on par with compared methods. However, these improvements need to be seen in the light of the difficulties of objectively evaluating and comparing medical segmentation and registration algorithms. Choice of similarity metrics, evaluation data and tuning parameters may greatly impact the results, and re-running competing methods locally may lead to unfair experimental setups. Moreover, installing, tuning and running competing methods are time-consuming and implementations of current state-of-the-art may not be publicly accessible. Due to this, only a fraction of previously published methods are used for comparison, which of course is crippling for evaluations said to be meticulous.

Ideally, each method should be evaluated on unseen test data provided by a public benchmark. Moreover, the public benchmark should decide for relevant accuracy metrics. In this way, unbiased conclusions can be made directly with means of a public benchmark leaderboard, without needing to re-run compared methods locally. This procedure enables fair comparisons between competing methods, unfortunately, there is often no such benchmark available for the particular application at hand. In fact, this is the case for a majority of the thesis datasets.

Papers I, II and V validate the proposed methods on publicly accessible data, the Adni, Visceral and Hammers datasets. For the Adni and the Hammers dataset, the proposed methods are evaluated on the same test set as reported in the publications corresponding to the compared methods. Thus, there is no need to re-run the compared methods locally, and fair comparisons to these particular methods can be made directly as described above. Unfortunately, there are no public leaderboards summarizing all previous attempts on these particular datasets. For the Visceral dataset, evaluation is done by splitting the provided training data into smaller sets and by running public or in-house versions of the competing methods. Hyperparameters are chosen following the recommendations in the corresponding papers. Thus, the evaluation suffer from an unfair experimental setup, since the competing methods are not tuned with respect to the dataset at hand.

Paper III does validate the proposed method on test data from a public benchmark with favorable results. However, this benchmark dataset is not a medical dataset, which makes it hard to draw any conclusions regarding the potential for medical applications. The medical datasets in the paper, the Scapis and the Echo datasets, are unfortunately not public. That is, there are no previous results on these particular datasets to compare with. Of course, this makes it hard to draw any objective conclusions. The same holds for paper IV, only the Scapis dataset is considered due to the very task-specific objective (delineating the pericardium). Unfortunately, neither datasets nor implemented versions of previously proposed methods for pericardium segmentation are (as of date) publicly available. These difficulties highlights the advantages of general-purpose algorithms independent of application and modality: only considering one, or a few, specific tasks makes comparisons highly inconvenient in the absence of benchmark databases.

4.1.2 Running times and scalability

One paper out of five, paper V, addresses running time (and scalability) as an explicit research objective. The paper does succeed in speeding up parts of the multi-atlas framework, however, some time-consuming steps are heavily overlooked (such as image warping and non-rigid registration). Moreover, the paper does not report the running times for all compared methods. Some of the papers (I, II and IV) acknowledge the running time as an issue, but lack in evaluation. Paper II and IV merely report the running times for the proposed methods. Paper I also reports running times for some, but not all, of the compared methods. Ideally, running time comparisons should be carried out on the same hardware for all compared methods. Again, such fair comparisons would be helped by a framework for benchmark comparisons as discussed above.

4.1.3 Limited datasets and generalizability

Limited access to labelled data is a common issue in medical image analysis, and also in a majority of the included thesis papers. Standard solutions, such as cross-validation and data augmentation, are included in all papers. For a majority of the papers, the lack of labelled data makes it hard to draw conclusions regarding generalizability, since this objective should be evaluated on large, and diverse, datasets. Again, publicly available benchmark datasets including a large amount of diverse cases would definitely simplify evaluation of properties such as generalizability and robustness.

Two papers, III and IV, address the lack of labelled data explicitly. For paper IV, it remains somewhat unclear whether the proposed solution (rotation-invariant features) manages to tackle this issue, especially since the test set is very limited (10 subjects) and similar to the training set. Paper III evaluates the method on several different datasets, where one of the datasets (Scapis) includes 300 test images. This should indeed indicate that the proposed method is capable to generalize. However, paper I does present the most convincing evaluation when it comes to generalizability. In this paper, the proposed method is successfully evaluated on 111 test subjects from a completely different dataset than the training set, which should indicate that the method is general and robust.

4.1.4 Anatomical and physiological plausibility

Two out of five of the papers address the qualitative segmentation shape as an research objective, paper II and III. Paper III evaluates the improvement with means of qualitative (visual) examples, however, the included evaluation metric (the Jaccard index) does not capture the qualitative shape very well. Paper II presents a, perhaps, more convincing evaluation of the qualitative shape with means of a more suitable evaluation metric (the Hausdorff distance) in addition to visual examples. However, qualitative shape is highly subjective and thus difficult to quantify. A thorough visual inspection by a medical expert would most surely benefit both evaluations.

In paper II, the shape prior did seem to impact the segmentation negatively for some cases, leading to over-regularized boundaries, which may infuse doubt regarding the generalizability. Additionally, one may dispute the choice of merely including the shape-regularized segmentation as input to a classifier, and not directly enforcing the refined solution to cohere with the shape prior. In retrospect, one could consider executing a comparison to a similar classifier merely trained on pure image features. Paper III proposes a method for incorporating shape regularization that should be more capable of generalizing, thanks to the max-margin loss and the end-to-end training. A discussed above, the evaluation does indeed indicate good generalization capabilities.

4.2 Future directions

4.2.1 Shape and learning for coronary artery segmentation

Buildup of vascular plaque in the coronary arteries, the blood vessels that provide oxygenated blood to the heart, is a biomarker for myocardial infarction ("heart attack"). Vascular plaques cause stenosis, a narrowing of the blood vessel, that can be detected and quantified by comparing the width of the lumen and the vessel wall of the coronary arteries. Automatic segmentation of coronary arteries, including lumen, wall and plaque delineations, can help assessing the risk of myocardial infarctions. Currently, 600 cardiac CTA images from the Scapis dataset are manually annotated for coronary artery segmentations. The manual annotations include lumen and vessel wall of the main coronary arteries (vessels large enough to be deemed medically relevant) as well as plaques and stenoses. The aim is to train a deep model able to detect and classify plaques and stenoses with an accuracy comparable to an expert radiologist.

The idea of using a deep network for analysing the coronary artery tree is not new, see [138, 139]. However, these previous attempts neither provide full segmentations nor enforce necessary shape constraints. The anatomy of coronary arteries cannot vary in every possible way, that is, there are a number of anatomical constraints that an automatic software should obey. The lumen, and plaques if present, always lie inside the vessel wall, the cross-section of the vessel wall has a convex shape and the coronary arteries grow like a tree from the aorta. Conditional random field models could be one way of posing such shape constraints on the output segmentation while still generalizing well to unseen data. Previous work has proposed CRF models enforcing star-shaped and convex shapes [140–142], relative position of multiple regions [143,144] and tree structures [145]. The tubular tree structure could also be enforced by other shape modelling techniques, such as Active Shape Models (ASMs) as in [146], or shortest geodesic path trees as in [147]. Enabling end-to-end training of deep networks coupled with shape models enforcing these necessary geometric priors is yet to be done, and will surely boost the qualitative performance of a deep segmentation algorithm for coronary artery segmentation.

4.2.2 Generalizable deep models for echo assessment

Echocardiography (cardiac ultrasound, "echo") is a widely used medical imaging technique for heart function assessment. However, it takes several years to train a physician to become a senior specialist in echocardiography assessment. An automatic echocardiography assessment software could be of great use in the clinical work: it could work as a as a gate keeper to sort out the less complex cases, it could reduce the burden of physicians and it could enable echocardiography assessments at clinic sites with no senior echocardiography specialist. A unique dataset at the Sahlgrenska university hospital, including over 90000 echo examinations with manual annotations of the size, function and disease status of the heart ventricles and atriums as well the heart valves, should allow for training a deep model able to automatically assess an echo examination.

Using existing clinical data for training and evaluating a deep model able to perform a diverse set of tasks is of course a challenging problem. The examinations are acquired during different years, by different physicians, from different view points and with means of different scanner types. The assessments vary from volume estimation to valve function classification. Thus, the model must generalize between tasks and between domains. Further, the model should be robust to bad quality examinations, perhaps with missing images or performed by inexperienced physicians. Finally, the memory footprint of each examination and the very size of the dataset put demands on running times and hardware. Developing a deep model that successfully masters these requirements on robustness, generalizability and scalability would open up for other projects using already existing clinical data, instead of relying on new time-consuming manual annotation.

Bibliography

[1] P. Suetens, *Fundamentals of medical imaging.* Cambridge University Press, 2017.

[2] R. H. Taylor, A. Menciassi, G. Fichtinger, and P. Dario, "Medical robotics and computer-integrated surgery," in *Handbook of Robotics.* Springer, 2008, pp. 1199–1222.

[3] D. L. Pham, C. Xu, and J. L. Prince, "Current methods in medical image segmentation 1," *Annual Review of Biomedical Engineering*, vol. 2, no. 1, pp. 315–337, 2000.

[4] G. Bergström *et al.*, "The Swedish CArdioPulmonary bioImage study: Objectives and design," *Journal of Internal Medicine*, vol. 278, no. 6, pp. 645–659, 2015.

[5] O. A. Jimenez del Toro *et al.*, "Cloud-based evaluation of anatomical structure segmentation and landmark detection algorithms: VISCERAL anatomy benchmarks," *IEEE Transactions on Medical Imaging*, vol. 35, no. 11, pp. 2459–2475, 2016.

[6] A. Hammers *et al.*, "Three-dimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe," *Human Brain Mapping*, vol. 19, no. 4, pp. 224–247, 2003.

[7] I. S. Gousias *et al.*, "Automatic segmentation of brain MRIs of 2-year-olds into 83 regions of interest," *NeuroImage*, vol. 40, no. 2, pp. 672–684, 2008.

[8] A. Sotiras, C. Davatzikos, and N. Paragios, "Deformable medical image registration: A survey," *IEEE Transactions on Medical Imaging*, vol. 32, no. 7, pp. 1153–1190, 2013.

[9] J. A. Maintz and M. A. Viergever, "A survey of medical image registration," *Medical Image Analysis*, vol. 2, no. 1, pp. 1–36, 1998.

[10] J. E. Iglesias and M. R. Sabuncu, "Multi-atlas segmentation of biomedical images: A survey," *Medical Image Analysis*, vol. 24, no. 1, pp. 205–219, 2015.

[11] O. Hansson *et al.*, "Tau pathology distribution in Alzheimer's disease corresponds differentially to cognition-relevant functional brain networks," *Frontiers in Neuroscience*, vol. 11, p. 167, 2017.

[12] W. E. Brant and C. A. Helms, *Fundamentals of diagnostic radiology*. Lippincott Williams & Wilkins, 2012.

[13] F. Khalifa, G. M. Beache, G. Gimel'farb, J. S. Suri, and A. S. El-Baz, "State-of-the-art medical image registration methodologies: A survey," in *Multi Modality State-of-the-art Medical Image Segmentation and Registration Methodologies*. Springer, 2011, pp. 235–280.

[14] Y. Ou, A. Sotiras, N. Paragios, and C. Davatzikos, "DRAMMS: Deformable registration via attribute matching and mutual-saliency weighting," *Medical Image Analysis*, vol. 15, no. 4, pp. 622–639, 2011.

[15] S. Ourselin, A. Roche, G. Subsol, X. Pennec, and N. Ayache, "Reconstructing a 3D structure from serial histological sections," *Image and Vision Computing*, vol. 19, no. 1, pp. 25–31, 2001.

[16] S. Ourselin, R. Stefanescu, and X. Pennec, "Robust registration of multimodal images: Towards real-time clinical applications," in *Conference on Medical Image Computing and Computer-Assisted Intervention*, 2002, pp. 140–147.

[17] T. Vercauteren, X. Pennec, A. Perchant, and N. Ayache, "Diffeomorphic demons: Efficient non-parametric image registration," *NeuroImage*, vol. 45, no. 1, pp. S61–S72, 2009.

[18] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. Pluim, "Elastix: A toolbox for intensity-based medical image registration," *IEEE Transactions on Medical Imaging*, vol. 29, no. 1, pp. 196–205, 2010.

[19] B. B. Avants, N. J. Tustison, M. Stauffer, G. Song, B. Wu, and J. C. Gee, "The Insight ToolKit image registration framework," *Frontiers in Neuroinformatics*, vol. 8, 2014.

[20] J. P. Pluim, J. A. Maintz, and M. A. Viergever, "Mutual-information-based registration of medical images: A survey," *IEEE Transactions on Medical Imaging*, vol. 22, no. 8, pp. 986–1004, 2003.

[21] F. Maes, D. Vandermeulen, and P. Suetens, "Comparative evaluation of multiresolution optimization strategies for multimodality image registration by maximization of mutual information," *Medical Image Analysis*, vol. 3, no. 4, pp. 373–386, 1999.

[22] S. Klein, M. Staring, and J. P. Pluim, "Evaluation of optimization methods for nonrigid medical image registration using mutual information and B-splines," *IEEE Transactions on Image Processing*, vol. 16, no. 12, pp. 2879–2890, 2007.

[23] 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.

[24] M. F. Beg, M. I. Miller, A. Trouvé, and L. Younes, "Computing large deformation metric mappings via geodesic flows of diffeomorphisms," *International Journal of Computer Vision*, vol. 61, no. 2, pp. 139–157, 2005.

[25] B. B. Avants, C. L. Epstein, M. Grossman, and J. C. 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.

[26] F. L. Bookstein, "Principal warps: Thin-plate splines and the decomposition of deformations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 6, pp. 567–585, 1989.

[27] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. Hill, M. O. Leach, and D. J. Hawkes, "Nonrigid registration using free-form deformations: Application to breast MR images," *IEEE Transactions on Medical Imaging*, vol. 18, no. 8, pp. 712–721, 1999.

[28] L. Svärm, O. Enqvist, F. Kahl, and M. Oskarsson, "Improving robustness for inter-subject medical image registration using a feature-based approach," in *International Symposium on Biomedical Imaging*, 2015, pp. 824–828.

[29] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[30] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[31] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv preprint arXiv:1312.6229*, 2013.

[32] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. Moreno-Noguer, "Discriminative learning of deep convolutional feature point descriptors," in *IEEE International Conference on Computer Vision*, 2015, pp. 118–126.

[33] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua, "Lift: Learned invariant feature transform," in *European Conference on Computer Vision*. Springer, 2016, pp. 467–483.

[34] P. Fischer, A. Dosovitskiy, and T. Brox, "Descriptor matching with convolutional neural networks: A comparison to SIFT," *arXiv preprint arXiv:1405.5769*, 2014.

[35] X. Han, T. Leung, Y. Jia, R. Sukthankar, and A. C. Berg, "Matchnet: Unifying feature and metric learning for patch-based matching," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3279–3286.

[36] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[37] S. Lee, G. Wolberg, and S. Y. Shin, "Scattered data interpolation with multilevel B-splines," *IEEE Transactions on Visualization and Computer Graphics*, vol. 3, no. 3, pp. 228–244, 1997.

[38] H. Chui and A. Rangarajan, "A new point matching algorithm for non-rigid registration," *Computer Vision and Image Understanding*, vol. 89, no. 2, pp. 114–141, 2003.

[39] 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, no. 2, pp. 239–256, 1992.

[40] C. V. Stewart, C.-L. Tsai, and B. Roysam, "The dual-bootstrap iterative closest point algorithm with application to retinal image registration," *IEEE Transactions on Medical Imaging*, vol. 22, no. 11, pp. 1379–1394, 2003.

[41] T. Rohlfing, R. Brandt, R. Menzel, and C. R. Maurer, "Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains," *NeuroImage*, vol. 21, no. 4, pp. 1428–1442, 2004.

[42] A. Klein, B. Mensh, S. Ghosh, J. Tourville, and J. Hirsch, "Mindboggle: Automated brain labeling with multiple atlases," *BMC Medical Imaging*, vol. 5, no. 1, p. 7, 2005.

[43] R. A. Heckemann, J. V. Hajnal, P. Aljabar, D. Rueckert, and A. Hammers, "Automatic anatomical brain MRI segmentation combining label propagation and decision fusion," *NeuroImage*, vol. 33, no. 1, pp. 115–126, 2006.

[44] P. Aljabar, R. A. Heckemann, A. Hammers, J. V. Hajnal, and D. Rueckert, "Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy," *NeuroImage*, vol. 46, no. 3, pp. 726–738, 2009.

[45] T. R. Langerak, U. A. van der Heide, A. N. Kotte, M. A. Viergever, M. Van Vulpen, and J. P. Pluim, "Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (SIMPLE)," *IEEE Transactions on Medical Imaging*, vol. 29, no. 12, pp. 2000–2008, 2010.

[46] X. Artaechevarria, A. Muñoz-Barrutia, and C. Ortiz-de Solorzano, "Efficient classifier generation and weighted voting for atlas-based segmentation: Two small steps faster and closer to the combination oracle," *SPIE Medical Imaging: Image Processing*, vol. 6914, no. 3, pp. 69 141W–1, 2008.

[47] X. Artaechevarria, A. Munoz-Barrutia, and C. Ortiz-de Solórzano, "Combination strategies in multi-atlas image segmentation: Application to brain MR data," *IEEE Transactions on Medical Imaging*, vol. 28, no. 8, pp. 1266–1277, 2009.

[48] H. Wang, J. W. Suh, S. R. Das, J. B. Pluta, C. Craige, and P. A. Yushkevich, "Multi-atlas segmentation with joint label fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 3, pp. 611–623, 2013.

[49] R. Wolz, C. Chu, K. Misawa, M. Fujiwara, K. Mori, and D. Rueckert, "Automated abdominal multi-organ segmentation with subject-specific atlas generation," *IEEE Transactions on Medical Imaging*, vol. 32, no. 9, pp. 1723–1730, 2013.

[50] G. Wu, Q. Wang, D. Zhang, F. Nie, H. Huang, and D. Shen, "A generative probability model of joint label fusion for multi-atlas based brain segmentation," *Medical Image Analysis*, vol. 18, no. 6, pp. 881–890, 2014.

[51] Y. Song, G. Wu, Q. Sun, K. Bahrami, C. Li, and D. Shen, "Progressive label fusion framework for multi-atlas segmentation by dictionary evolution," in *Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 190–197.

[52] T. Rohlfing, D. B. Russakoff, and C. R. Maurer, "Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation," *IEEE Transactions on Medical Imaging*, vol. 23, no. 8, pp. 983–994, 2004.

[53] S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image

segmentation," *IEEE Transactions on Medical Imaging*, vol. 23, no. 7, pp. 903–921, 2004.

[54] Z. Xu *et al.*, "Efficient multi-atlas abdominal segmentation on clinically acquired CT with SIMPLE context learning," *Medical Image Analysis*, vol. 24, no. 1, pp. 18–27, 2015.

[55] M. R. Sabuncu, B. T. Yeo, K. Van Leemput, B. Fischl, and P. Golland, "A generative model for image segmentation based on label fusion," *IEEE Transactions on Medical Imaging*, vol. 29, no. 10, pp. 1714–1729, 2010.

[56] H. Yang, J. Sun, H. Li, L. Wang, and Z. Xu, "Deep fusion net for multi-atlas segmentation: Application to cardiac MR images," in *Conference on Medical Image Computing and Computer-Assisted Intervention*, 2016, pp. 521–528.

[57] A. Blake, P. Kohli, and C. Rother, *Markov random fields for vision and image processing*. MIT Press, 2011.

[58] C. Sutton and A. McCallum, "An introduction to conditional random fields," *Foundations and Trends ® in Computer Graphics and Vision*, vol. 4, no. 4, pp. 267–373, 2012.

[59] C. Wang, N. Komodakis, and N. Paragios, "Markov random field modeling, inference & learning in computer vision & image understanding: A survey," *Computer Vision and Image Understanding*, vol. 117, no. 11, pp. 1610–1627, 2013.

[60] Y. Boykov and M.-P. Jolly, "Interactive organ segmentation using graph cuts," in *Conference on Medical Image Computing and Computer-Assisted Intervention*, 2000, pp. 276–286.

[61] Y. Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in ND images," in *IEEE International Conference on Computer Vision*, vol. 1, 2001, pp. 105–112.

[62] S. Esneault, C. Lafon, and J.-L. Dillenseger, "Liver vessels segmentation using a hybrid geometrical moments/graph cuts method," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 2, pp. 276–283, 2009.

[63] X. Chen, J. K. Udupa, U. Bagci, Y. Zhuge, and J. Yao, "Medical image segmentation by combining graph cuts and oriented active appearance models," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 2035–2046, 2012.

[64] X. Chen and L. Pan, "A survey of graph cuts/graph search based medical image segmentation," *IEEE reviews in biomedical engineering*, vol. 11, pp. 112–124, 2018.

[65] C. Platero and M. C. Tobar, "A label fusion method using conditional random fields with higher-order potentials: Application to hippocampal segmentation," *Artificial Intelligence in Medicine*, vol. 64, no. 2, pp. 117–129, 2015.

[66] V. Kolmogorov and Y. Boykov, "What metrics can be approximated by geocuts, or global optimization of length/area and flux," in *IEEE International Conference on Computer Vision*, vol. 1, 2005, pp. 564–571.

[67] A. K. Sinop and L. Grady, "A seeded image segmentation framework unifying graph cuts and random walker which yields a new algorithm," in *IEEE International Conference on Computer Vision*, 2007, pp. 1–8.

[68] V. Kolmogorov and R. Zabin, "What energy functions can be minimized via graph cuts?" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 2, pp. 147–159, 2004.

[69] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 11, pp. 1222–1239, 2001.

[70] M. de Bruijne, "Machine learning approaches in medical image analysis: From detection to diagnosis," *Medical Image Analysis*, vol. 33, pp. 94–97, 2016.

[71] T. K. Ho, "Random decision forests," in *Proceedings of 3rd international conference on document analysis and recognition*, vol. 1, 1995, pp. 278–282.

[72] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[73] E. Konukoglu *et al.*, "Robust linear registration of CT images using random regression forests," in *Medical Imaging 2011: Image Processing*, vol. 7962, 2011, p. 79621X.

[74] P.-H. Conze, F. Tilquin, V. Noblet, F. Rousseau, F. Heitz, and P. Pessaux, "Hierarchical multi-scale supervoxel matching using random forests for automatic semi-dense abdominal image registration," in *International Symposium on Biomedical Imaging*, 2017, pp. 490–493.

[75] F. Kanavati *et al.*, "Supervoxel classification forests for estimating pairwise image correspondences," *Pattern Recognition*, vol. 63, pp. 561–569, 2017.

[76] T. Lotfi, L. Tang, S. Andrews, and G. Hamarneh, "Improving probabilistic image registration via reinforcement learning and uncertainty evaluation,"

in *International Workshop on Machine Learning in Medical Imaging*, 2013, pp. 187–194.

[77] D. Han, Y. Gao, G. Wu, P.-T. Yap, and D. Shen, "Robust anatomical landmark detection for MR brain image registration," in *Conference on Medical Image Computing and Computer-Assisted Intervention*, 2014, pp. 186–193.

[78] C. Lindner, S. Thiagarajah, J. M. Wilkinson, G. A. Wallis, T. F. Cootes, and arcOGEN Consortium, "Fully automatic segmentation of the proximal femur using random forest regression voting," *IEEE Transactions on Medical Imaging*, vol. 32, no. 8, pp. 1462–1472, 2013.

[79] D. Mahapatra, "Analyzing training information from random forests for improved image segmentation," *IEEE Transactions on Image Processing*, vol. 23, no. 4, pp. 1504–1512, 2014.

[80] D. Zikic *et al.*, "Decision forests for tissue-specific segmentation of high-grade gliomas in multi-channel MR," in *Conference on Medical Image Computing and Computer-Assisted Intervention*, 2012, pp. 369–376.

[81] M. Yaqub, M. K. Javaid, C. Cooper, and J. A. Noble, "Investigation of the role of feature selection and weighted voting in random forests for 3-D volumetric segmentation," *IEEE Transactions on Medical Imaging*, vol. 33, no. 2, pp. 258–271, 2013.

[82] N. J. Tustison *et al.*, "Optimal symmetric multimodal templates and concatenated random forests for supervised brain tumor segmentation (simplified) with ANTsR," *Neuroinformatics*, vol. 13, no. 2, pp. 209–225, 2015.

[83] A. Montillo, J. Shotton, J. Winn, J. E. Iglesias, D. Metaxas, and A. Criminisi, "Entangled decision forests and their application for semantic segmentation of CT images," in *Biennial International Conference on Information Processing in Medical Imaging*, 2011, pp. 184–196.

[84] B. Glocker, O. Pauly, E. Konukoglu, and A. Criminisi, "Joint classification-regression forests for spatially structured multi-object segmentation," in *European Conference on Computer Vision*, 2012, pp. 870–881.

[85] Y. Gao, Y. Shao, J. Lian, A. Z. Wang, R. C. Chen, and D. Shen, "Accurate segmentation of CT male pelvic organs via regression-based deformable models and multi-task random forests," *IEEE Transactions on Medical Imaging*, vol. 35, no. 6, pp. 1532–1543, 2016.

[86] R. Cuingnet, R. Prevost, D. Lesage, L. D. Cohen, B. Mory, and R. Ardon, "Automatic detection and segmentation of kidneys in 3D CT images using

random forests," in *Conference on Medical Image Computing and Computer-Assisted Intervention*, 2012, pp. 66–74.

[87] M. P. Heinrich and M. Blendowski, "Multi-organ segmentation using vantage point forests and binary context features," in *Conference on Medical Image Computing and Computer-Assisted Intervention*, 2016, pp. 598–606.

[88] C. Jin *et al.*, "3D fast automatic segmentation of kidney based on modified AAM and random forest," *IEEE Transactions on Medical Imaging*, vol. 35, no. 6, pp. 1395–1407, 2016.

[89] J. Maiora, B. Ayerdi, and M. Graña, "Random forest active learning for AAA thrombus segmentation in computed tomography angiography images," *Neurocomputing*, vol. 126, pp. 71–77, 2014.

[90] A. Mansoor *et al.*, "A generic approach to pathological lung segmentation," *IEEE Transactions on Medical Imaging*, vol. 33, no. 12, pp. 2293–2310, 2014.

[91] K. Fukushima, "Neural network model for a mechanism of pattern recognition unaffected by shift in position- neocognitron," *Electron. & Commun. Japan*, vol. 62, no. 10, pp. 11–18, 1979.

[92] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[93] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587.

[94] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[95] D. Ciresan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Deep neural networks segment neuronal membranes in electron microscopy images," in *Advances in Neural Information Processing Systems*, 2012, pp. 2843–2851.

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

[97] A. Prasoon, K. Petersen, C. Igel, F. Lauze, E. Dam, and M. Nielsen, "Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network," in *Conference on Medical Image Computing and Computer-Assisted Intervention*, 2013, pp. 246–253.

[98] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)*, 2016, pp. 565–571.

[99] H. R. Roth *et al.*, "Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation," in *Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 556–564.

[100] Q. Dou, H. Chen, Y. Jin, L. Yu, J. Qin, and P.-A. Heng, "3D deeply supervised network for automatic liver segmentation from CT volumes," in *Conference on Medical Image Computing and Computer-Assisted Intervention*, 2016, pp. 149–157.

[101] R. Korez, B. Likar, F. Pernuš, and T. Vrtovec, "Model-based segmentation of vertebral bodies from MR images with 3D CNNs," in *Conference on Medical Image Computing and Computer-Assisted Intervention*, 2016, pp. 433–441.

[102] P. V. Tran, "A fully convolutional neural network for cardiac segmentation in short-axis MRI," *arXiv preprint arXiv:1604.00494*, 2016.

[103] W. Zhang *et al.*, "Deep convolutional neural networks for multi-modality isointense infant brain image segmentation," *NeuroImage*, vol. 108, pp. 214–224, 2015.

[104] T. Brosch, L. Y. Tang, Y. Yoo, D. K. Li, A. Traboulsee, and R. Tam, "Deep 3D convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1229–1239, 2016.

[105] P. Moeskops, M. A. Viergever, A. M. Mendrik, L. S. de Vries, M. J. Benders, and I. Išgum, "Automatic segmentation of MR brain images with a convolutional neural network," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1252–1261, 2016.

[106] M. Havaei *et al.*, "Brain tumor segmentation with deep neural networks," *Medical Image Analysis*, vol. 35, pp. 18–31, 2017.

[107] G. Wu, M. Kim, Q. Wang, Y. Gao, S. Liao, and D. Shen, "Unsupervised deep feature learning for deformable registration of MR brain images," in *Conference on Medical Image Computing and Computer-Assisted Intervention*, 2013, pp. 649–656.

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

[109] 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.

[110] A. V. Dalca, G. Balakrishnan, J. Guttag, and M. R. Sabuncu, "Unsupervised learning for fast probabilistic diffeomorphic registration," in *Conference on Medical Image Computing and Computer-Assisted Intervention*, 2018, pp. 729–738.

[111] 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, no. 8, pp. 1788–1800, 2019.

[112] K. A. Eppenhof, M. W. Lafarge, P. Moeskops, M. Veta, and J. P. Pluim, "Deformable image registration using convolutional neural networks," in *Medical Imaging 2018: Image Processing*, vol. 10574, 2018, p. 105740S.

[113] H. Sokooti, B. de Vos, F. Berendsen, B. P. Lelieveldt, I. Išgum, and M. Staring, "Nonrigid image registration using multi-scale 3D convolutional neural networks," in *Conference on Medical Image Computing and Computer-Assisted Intervention*, 2017, pp. 232–239.

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

[115] M.-M. Rohé, M. Datar, T. Heimann, M. Sermesant, and X. Pennec, "SVF-Net: Learning deformable image registration using shape matching," in *Conference on Medical Image Computing and Computer-Assisted Intervention*, 2017, pp. 266–274.

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

[117] Y. Hu *et al.*, "Label-driven weakly-supervised learning for multimodal deformable image registration," in *International Symposium on Biomedical Imaging*, 2018, pp. 1070–1074.

[118] Z. Wu, C. Shen, and A. Van Den Hengel, "Wider or deeper: Revisiting the resnet model for visual recognition," *Pattern Recognition*, vol. 90, pp. 119–133, 2019.

[119] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[120] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *International Conference on Machine Learning*, 2010, pp. 807–814.

[121] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *International Conference on Machine Learning*, vol. 30, no. 1, 2013, p. 3.

[122] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *IEEE International Conference on Computer Vision*, 2015, pp. 1026–1034.

[123] P. Ramachandran, B. Zoph, and Q. V. Le, "Swish: A self-gated activation function," *arXiv preprint arXiv:1710.05941*, vol. 7, 2017.

[124] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[125] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *IEEE International Conference on Computer Vision*, 2015, pp. 1520–1528.

[126] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, 2015, pp. 448–456.

[127] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *International Conference on Machine Learning*, 2013, pp. 1139–1147.

[128] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, vol. 12, no. Jul, pp. 2121–2159, 2011.

[129] M. D. Zeiler, "ADADELTA: An adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.

[130] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[131] T. Dozat, "Incorporating Nesterov momentum into Adam," 2016, Available from: http://cs229.stanford.edu/proj2015/054_report.pdf.

[132] Y. LeCun *et al.*, "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989.

[133] ——, "Handwritten digit recognition with a back-propagation network," in *Advances in Neural Information Processing Systems*, 1990, pp. 396–404.

[134] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[135] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[136] L. Prechelt, "Early stopping – but when?" in *Neural Networks: Tricks of the Trade*. Springer, 2012, pp. 53–67.

[137] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[138] M. Zreik, R. W. van Hamersvelt, J. M. Wolterink, T. Leiner, M. A. Viergever, and I. Išgum, "A recurrent CNN for automatic detection and classification of coronary artery plaque and stenosis in coronary CT angiography," *IEEE Transactions on Medical Imaging*, 2018.

[139] J. M. Wolterink, R. W. van Hamersvelt, M. A. Viergever, T. Leiner, and I. Išgum, "Coronary artery centerline extraction in cardiac CT angiography using a CNN-based orientation classifier," *Medical Image Analysis*, vol. 51, pp. 46–60, 2019.

[140] O. Veksler, "Star shape prior for graph-cut image segmentation," in *European Conference on Computer Vision*, 2008, pp. 454–467.

[141] V. Gulshan, C. Rother, A. Criminisi, A. Blake, and A. Zisserman, "Geodesic star convexity for interactive image segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3129–3136.

[142] L. Gorelick, O. Veksler, Y. Boykov, and C. Nieuwenhuis, "Convexity shape prior for segmentation," in *European Conference on Computer Vision*, 2014, pp. 675–690.

[143] A. Delong and Y. Boykov, "Globally optimal segmentation of multi-region objects," in *IEEE International Conference on Computer Vision*, 2009, pp. 285–292.

[144] J. Ulén, P. Strandmark, and F. Kahl, "An efficient optimization framework for multi-region segmentation based on Lagrangian duality," *IEEE Transactions on Medical Imaging*, vol. 32, no. 2, pp. 178–188, 2013.

[145] C. Bauer, T. Pock, E. Sorantin, H. Bischof, and R. Beichel, "Segmentation of interwoven 3D tubular tree structures utilizing shape priors and graph cuts," *Medical Image Analysis*, vol. 14, no. 2, pp. 172–184, 2010.

[146] M. de Bruijne, B. van Ginneken, M. A. Viergever, and W. J. Niessen, "Adapting active shape models for 3D segmentation of tubular structures in medical images," in *Biennial International Conference on Information Processing in Medical Imaging*, 2003, pp. 136–147.

[147] J. Stuhmer, P. Schroder, and D. Cremers, "Tree shape priors with connectivity constraints using convex relaxation on general graphs," in *IEEE International Conference on Computer Vision*, 2013, pp. 2336–2343.