Object Segmentation by Fitting Statistical Shape Models

A Kernel-Based Approach with Application to Wisdom Tooth Segmentation from CBCT Images

Inauguraldissertation

zur
Erlangung der Würde eines Doktors der Philosophie
vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von

Christoph Jud
aus Schänis St. Gallen

Basel, 2014

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät

auf Antrag von

Prof. Dr. Thomas Vetter, Universität Basel, Dissertationsleiter
Prof. Dr. Volker Roth, Universität Basel, Korreferent

Basel, den 24.6.2014

Prof. Dr. Jörg Schibler

Namensnennung-Keine kommerzielle Nutzung-Keine Bearbeitung 3.0 Schweiz
(CC BY-NC-ND 3.0 CH)

Sie dürfen: **Teilen** — den Inhalt kopieren, verbreiten und zugänglich machen

Unter den folgenden Bedingungen:

Namensnennung — Sie müssen den Namen des Autors/Rechteinhabers in der von ihm festgelegten Weise nennen.

Keine kommerzielle Nutzung — Sie dürfen diesen Inhalt nicht für kommerzielle Zwecke nutzen.

Keine Bearbeitung erlaubt — Sie dürfen diesen Inhalt nicht bearbeiten, abwandeln oder in anderer Weise verändern.

Wobei gilt:

- **Verzichtserklärung —** Jede der vorgenannten Bedingungen kann **aufgehoben** werden, sofern Sie die ausdrückliche Einwilligung des Rechteinhabers dazu erhalten.
- **Public Domain (gemeinfreie oder nicht-schützbare Inhalte) —** Soweit das Werk, der Inhalt oder irgendein Teil davon zur Public Domain der jeweiligen Rechtsordnung gehört, wird dieser Status von der Lizenz in keiner Weise berührt.
- **Sonstige Rechte —** Die Lizenz hat keinerlei Einfluss auf die folgenden Rechte:
    - Die Rechte, die jedermann wegen der Schranken des Urheberrechts oder aufgrund gesetzlicher Erlaubnisse zustehen (in einigen Ländern als grundsätzliche Doktrin des **fair use** bekannt);
    - Die **Persönlichkeitsrechte** des Urhebers;
    - Rechte anderer Personen, entweder am Lizenzgegenstand selber oder bezüglich seiner Verwendung, zum Beispiel für **Werbung** oder Privatsphärenschutz.
- **Hinweis —** Bei jeder Nutzung oder Verbreitung müssen Sie anderen alle Lizenzbedingungen mitteilen, die für diesen Inhalt gelten. Am einfachsten ist es, an entsprechender Stelle einen Link auf diese Seite einzubinden.

Quelle: http://creativecommons.org/licenses/by-nc-nd/3.0/ch/ Datum: 12.11.2013

To my parents.

ABSTRACT

Image segmentation is an important and challenging task in medical image analysis. Especially from low-quality images, segmentation algorithms have to cope with misleading background clutter, insufficient object boundaries and noise in the image. Statistical shape models are a powerful tool to tackle these problems. However, their construction as well as their application for segmentation remain challenging. In this thesis, we focus on the wisdom-tooth shape and its segmentation from Cone Beam Computed Tomography images. The large shape variation leads to difficult registration problems and an often too restrictive shape model, while the challenging appearance of the wisdom tooth makes the model fitting difficult.

To tackle these problems, we follow on kernel-based approaches to registration and shape modeling. We introduce a kernel, which considers landmarks as an additional prior in image registration. This allows to locally improve the registration accuracy. We present a *Demons*-like registration method with an inhomogeneous regularization which allows to apply such a landmark kernel.

For modeling the shape variation, we construct a kernel comprising a generic smoothness and an empirical sample covariance. With this combined kernel, we increase the flexibility of the statistical shape model. We make use of a reproducing kernel Hilbert space framework for registration, where we apply this combined kernel as reproducing kernel. To make the approach computationally feasible, we perform a low-rank approximation of the specific kernel function.

Because of a heterogeneous appearance inside the wisdom tooth, fitting the statistical model to plain intensity images is difficult. We build a nonparametric appearance model, based on random forest regression, which abstracts the raw images to semantic probability maps. Hence, the misleading structures become semantic values, which greatly simplificates the shape model fitting.

PUBLICATIONS

The main ideas, figures as well as experiments have appeared previously in the following publications:

[1] Christoph Jud. A flexible kernel framework for non-rigid image registration. Master's thesis, Computer Science Department, University of Basel, 2010.

[2] Christoph Jud and Thomas Vetter. Geodesically damped shape models. In *Proceedings of Shape 2014, Shape Symposium on Statistical Shape Models and Applications*. SICAS, 2014.

[3] Christoph Jud and Thomas Vetter. Using object probabilities in deformable model fitting. In *22st International Conference on Pattern Recognition (ICPR)*. IEEE, 2014.

[4] Christoph Jud, Marcel Lüthi, Thomas Albrecht, Sandro Schönborn, and Thomas Vetter. Variational image registration using inhomogeneous regularization. *Journal of Mathematical Imaging and Vision*, pages 1–15, 2014. ISSN 0924-9907. doi: 10.1007/s10851-014-0497-0.

[5] Marcel Lüthi, Christoph Jud, and Thomas Vetter. Using landmarks as a deformation prior for hybrid image registration. In *Pattern Recognition*, pages 196–205. Springer, 2011.

[6] Marcel Lüthi, Christoph Jud, and Thomas Vetter. A unified approach to shape model fitting and non-rigid registration. In *Machine Learning in Medical Imaging*, pages 66–73. Springer, 2013.

Further, the dignity of the science itself seems to require that every possible means be explored for the solution of a problem so elegant and so celebrated.

— Carl Friedrich Gauss,
in Article 329, Disquisitiones Arithmeticae, 1801

ACKNOWLEDGMENTS

First, I offer my sincerest gratitude to my supervisor, Prof. Dr. Thomas Vetter, who has supported me throughout my thesis. Without losing sight of the overall picture he consequently and critically reflected my work, which allowed me to grow as a research scientist. Beside my advisor, I thank my thesis commitee Prof. Dr. Volker Roth and Prof. Dr. Marcus Grote for letting my defense be an enjoyable moment and for the brilliant questions and comments.

Many other people have contributed to this work. Without their advices, their encouragement and numerous interesting scientific discussions this thesis would not have been possible. I thank everyone who supported me on this way. A special thank goes to my colleague, Dr. Marcel Lüthi, for our uncountable enlightening discussions and encouragements for my work. Furthermore, I thank Dr. Thomas Albrecht, Dr. Sandro Schönborn, André Graf and Thomas Gerig for a great collaboration and making for me the work at GraVis an inspiring and enjoyable time.

I am gratefol to the following people for proof-reading and helpful comments: Dr. Marcel Lüthi, Dr. Sandro Schönborn, Thomas Gerig, Bernhard Egger and Carole Flury.

Last but not least, the greatest thank goes to my parents Hermine and Hanspeter Jud for their unconditional support, their patience and their understanding in these four years.

CONTENTS

I INTRODUCTION & RELATED WORK 1
1 INTRODUCTION 3
1.1 Shape Models 4
1.2 Registration 4
1.3 Appearance Models 5
1.4 Contribution 7
1.5 Related Work 9
1.6 Overview 13

II SHAPE MODELING & OBJECT SEGMENTATION 15
2 REGISTRATION 17
2.1 Problem Description 17
2.1.1 Regularized Minimization Functional 19
2.1.2 Hybrid Image Registration 20
2.1.3 Shape Representations 21
2.2 Variational Registration 22
2.2.1 Linear Operators 24
2.2.2 Neighborhood Preservation 25
2.2.3 Inhomogeneous Regularization 27
2.2.4 Incorporating Landmarks 28
2.3 Discrete Filtering Approach 30
2.3.1 Separable Filter Decomposition 31
2.3.2 Efficient Nonstationary Filtering 33
2.3.3 Multi-Resolution 35
2.4 RKHS Registration 36
2.4.1 Minimizing in the RKHS 37
2.4.2 Low-Rank GP Model 40
2.4.3 Landmark Posterior Model 44
3 STATISTICAL MODEL 49
3.1 Statistical Shape Model 49
3.1.1 Extended Statistical Model Kernels 53
3.2 Statistical Appearance Model 56
3.2.1 Random Forest Regression 59
3.2.2 Probabilistic Similarity Metric 60
3.2.3 Shape to image metric 64
4 EXPERIMENTAL RESULTS 67

4.1 Variational Registration Experiments 68
4.1.1 POPI Breathing Thorax Model 69
4.1.2 Patellar Surface Registration 75
4.2 Low-rank Gaussian Process Registration vs. Demons Registration 80
4.3 Wisdom Tooth Shape Model 81
4.3.1 Shape Model Bias 82
4.4 Wisdom Tooth Segmentation 84
4.4.1 Evaluation of Segmentation 86
4.4.2 Geodesically Damped Model 91
4.5 Discussion 94
5 CONCLUSIONS 97
5.1 Future Work 98

III APPENDIX 101
A KERNEL REGISTRATION FRAMEWORK 103
A.1 Translation-Invariant Kernel Functions 103
A.2 Kernel Compositions 104
A.3 Minimization in RKHS Registration 105
A.4 PCA and SVD 105
A.5 Statistical Model Kernel 106
A.6 Functional Derivatives 106
A.7 C2Circle 107
B TENSOR DECOMPOSITIONS 109
B.1 Alternating Least Squares Method 109
B.2 Tucker Decomposition 110
B.3 Matrix Products 110

BIBLIOGRAPHY 113

NOTATION

$\mathbb{R}$	Real numbers.
$\mathbb{R}^d$	d-dimensional Euclidean space.
$\mathbb{N}$	Natural numbers $\{1, 2, \dots\}$
Ω	Image domain.
Γ	Surface domain.
$\circ$	Composition of functions.
$*$	Convolution operation.
δ_x	Dirac delta at location x.
$\lvert \cdot \rvert$	Absolute value.
$\lVert \cdot \rVert$	Euclidean norm.
$\mathcal{H}$	Hilbert space.
$\lVert \cdot \rVert_k$	Reproducing kernel Hilber space norm.
$\mathbf{I}_n$	Identity matrix of dimension n.
A^*	Adjoint operator of A.
∇	Gradient operator.
Δ	Laplacian operator.
$\nabla\cdot$	Divergence operator.

Part I

INTRODUCTION & RELATED WORK

1 INTRODUCTION

Medical imaging techniques have become very important in preoperative surgery planning and risk assessments as well as in disease diagnostics. Specifically, their noninvasive nature makes them indispensable. On the basis of only images, obtained from a computed tomography scanner for instance, an overview over the patient's anatomy can be derived. However, in the 3D case, it is difficult to a-priori recognize or see the three dimensional structure of an organ or a bone by looking solely at slices of the image. Thus, having the possibility to look at a rendered 3D scene of the patient's principal anatomical structures is much more comfortable. To generate such a scene, the objects of interest have to be virtually segmented in the image. However, the manual slice-by-slice segmentation is a tedious task and can lead to tendinitis originating by an excessive usage of the computer mouse (see Figure 1). Further, it can be difficult to determine, where the border of the object is, since the 3D structure can only be guessed by seeing the stacked 2D image slices. Thus, computer added segmentation is crucial to reach fast and accurate segmentations.

3D Imaging

Since medical imaging techniques have improved and their handling has become more comfortable, the need for applying them in routine examination has increased in the past few years. Hence, techniques such as Cone Beam Computed Tomography (CBCT) became popular, because of their low cost and the use of only low radiation. Especially in dental medicine, CBCT devices are broadly used, because they allow for examinations where the patient is not exposed to an intensive radiation.

However, this comes along with a loss of image quality, which poses a challenge to the segmentation algorithms. The low radiation has the effect that the resulting images have a low contrast and a small signal to noise ratio. In Figure 2, a slice through an image depicts a typical jaw region. Between the tooth's roots and the jaw there is nearly no difference in intensity. Hence, it is difficult to distinguish between tooth structure (object) and surrounding jaw (background). Further, the heavy noise in the image disturbs the homogeneous appearance of the object of interest. Segmentation algorithms for this kind of images have to cope with insufficient object boundaries, misleading background clutter and a strong noise.

Low-quality Images

Figure 1: After segmentation of few wisdom teeth from CBCT images, I got a tendinitis in my right arm. Fortunately, I am equipped with another arm which lets me to use the computer mouse now with the left hand.

Wisdom Tooth Shape

1.1 SHAPE MODELS

Statistical Shape Prior

To reach a global consistency across the segmented object shape, despite the mentioned difficulties, the application of statistical shape models is well established. The main idea is to introduce a prior over shapes, which are expected as a segmentation result. This prior is usually built upon given sample shapes, which represent the class of target shapes as good as possible. Having modeled the variation of shape within this class, the segmentation can be quantified in terms of such a shape prior, in order to consider only *meaningful* shapes for segmentation. A challenge remains, if there are few samples available compared to the variation of the object class, since the shape prior could become too restrictive. We will address this problem and show opportunities to combine the sample based prior with generic smoothness priors.

1.2 REGISTRATION

The modeling of the variation of the target shape class requires that for each point on a shape, the corresponding point in the remaining other sample shapes is known. Thus, at each point, the distribution over the shape variation is assessable. To put it another way, a (smooth) transformation from a reference shape to a target shape is desired. The process, where such a transformation is sought, which gives rise to the point-to-point correspondence between the shapes, is referred to as registration.

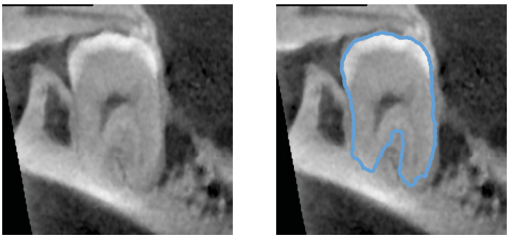


Figure 2: In the left figure, a slice through a CBCT image is shown. The right one shows the same slice but overlayed with the ground truth tooth outline (*blue*).

Correspondence Problem

The expressiveness of the final shape prior hinges on the quality of the registration, whose derivation is, however, an ill-posed problem. That means, a solution to the registration of two shapes, if it exists, is not unique, since the concept of correspondence between arbitrary shapes is ambiguous. Usually, smoothness heuristics for the transformation between the shapes are used to reduce this ambiguity. Moreover, in practice, it is important that the user is enabled to locally improve the registration. By providing *landmarks*, the user adds pairs of corresponding points, which influence the shape prior for a refined registration. In Figure 3, examples are shown, where the incorporation of the landmarks greatly improves the registration of wisdom tooth shapes. A main topic of this work is, how different smoothness assumptions as well as landmarks can be integrated into the registration.

1.3 APPEARANCE MODELS

In deformable shape model based segmentation, a deformation to a reference shape is sought, such that it matches the shape in the target image. The matching criterion depends not only on the shape prior but also on a similarity measure between the varied reference and the target shape. However, the plain intensity values are often not suited to be considered in the similarity measure, since they do not directly match structural facts. For instance, the roots and the jaw bone (see Figure 2) share equal pixel intensities but are different objects. Furthermore, the appearance inside the shape could be very heterogeneous which can violate the correspondence assumption across the object class. To overcome

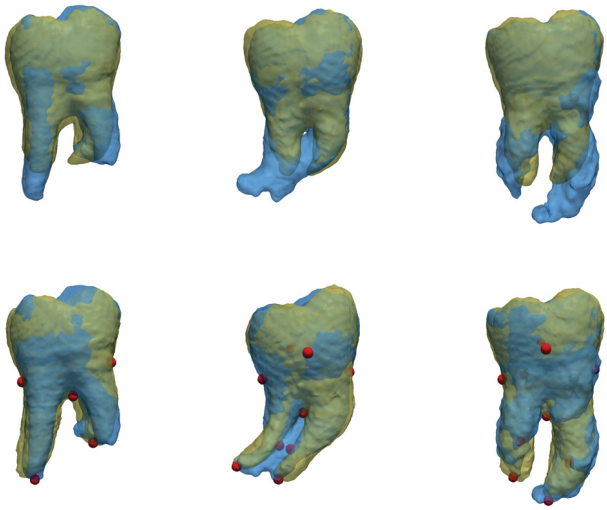


Figure 3: In this figure, the ground truth target shapes are colorized in *blue*, while the registered reference shapes are colorized in *yellow*. In the first row, no landmarks have been used for registration, which yields bad registration results. In these examples, this is mainly because of insufficient image features. In the bottom row, landmarks, visualized as *red* dots, have been incorporated applying the hybrid registration approach presented in this work. The resulting registrations (*yellow*) better match the target shapes (*blue*).

Abstraction of Low-level Cues

these issues, appearance models are in wide use. The representation of the image by an appearance model abstracts the physically obtained pixel intensities to semantic values. Ideally, such an abstraction allows to distinguish between objects with equal pixel intensities and is robust against the strong noise. In this work, we present a non-parametric appearance model, which is learned from sample data. In particular, we focus on models for object/background prediction to make the statistical shape model based segmentation more robust. In Figure 4, example object predictions are depicted, where on the right picture, beside the bright tooth structures, the dark inside of the tooth is predicted as object as well.

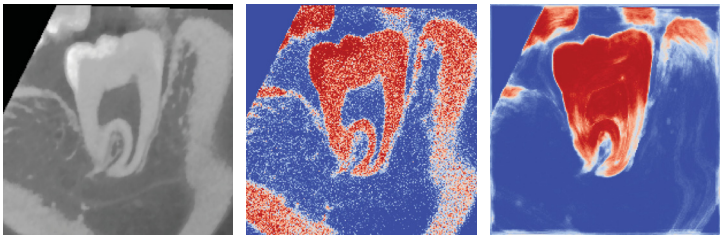


Figure 4: The figure shows a slice through a denoised CT image (left). Based on the pixel intensities, dentin and jaw as well as pulp and air cannot be distinguished. An appearance model of the tooth, based on solely the pixel intensities, leads to similar problems. In the middle, this issue is visualized, where the object probability of the tooth, based on an intensity based model is shown (*red*=1, *blue*=0). The pulp (dark inside of the tooth) cannot be predicted as object. However, using neighborhood dependent features, higher level structures can be learned. In the right object probability map, the jaw and air have a low, while the dentin and the pulp have a high object probability.

1.4 CONTRIBUTION

Our work has been inspired by the task of segmenting the wisdom tooth from a CBCT image for preoperative risk analysis. The low image quality motivates the use of statistical shape models. However, the variation of the wisdom tooth shape is large and difficult to model, while the number of sample images is relatively rare in our case. The consequence is, that the statistical model becomes too restrictive and does not generalize well. As such, in this work, we focus on kernel based approaches for implementing smoothness priors, to reach a framework, where priors can be flexibly changed and combined. For example, we construct kernels to reduce the model bias in order to make the statistical model more flexible.

Two approaches for registration are considered, where the first is based on the variational registration approach [14, 19, 93, 94][1], where a

1 These approaches are also known as *Demons*-like registration methods, where the Gaussian filtering approximates the diffusion regularization.

partial differential equation is solved, and in the second one, a solution is sought directly in a reproducing kernel Hilbert space (RKHS).

REGISTRATION AND SHAPE MODELS

- **Inhomogeneous Regularization**
  We generalize the variational registration approach, to spatially varying regularization (*i.e.* not translation-invariant). Thus, in our approach, a kernel which incorporates landmarks can be integrated as prior, which enables a hybrid landmark and image registration.
- **Efficient Nonstationary Filtering**
  The variational approach is based on an iterative convolution scheme. Hence, for the inhomogeneous regularization an efficient nonstationary filtering has to be implemented. Using the Tucker decomposition [96], we compress the local filter kernels, which leads to an efficiently nonstationary filter method. For stationary kernels, we apply low-rank approximation of the filter kernel to achieve fast separable filtering.
- **Incorporating Landmarks in RKHS Registration**
  We show, how the same hybrid landmark prior can be integrated in the very general registration method of Schölkopf et al. [87], where a solution is directly sought in the corresponding RKHS.
- **Low-rank Approximation of Kernels**
  To take advantage of this flexible registration framework, even for large 3D images, we approximate an orthogonal basis of the corresponding kernel function using the Nyström approximation.
- **A Framework for Statistical Model Bias Reduction**
  This formulation leads to a unified framework for registration and statistical shape model fitting. With the combination of both, a statistical model and a generic smoothness prior, we reduce the model bias and thus relax the restrictions of the shape model, to reach better fitting performance.
- **Geodesically Damped Kernels**
  In addition, we construct a geodesic surface prior. This prior damps correlations between parts on the surface of the shape, which are geodesically wide apart. With this prior, the model is equipped with a localized property which reduces the bias of statistical surface models as well.

APPEARANCE MODEL Moreover, for the statistical shape model based segmentation, we present a combined approach of a parametric shape and a nonparametric appearance model.

- **Random Forest Regression as Appearance Model**
  The appearance model is based on random forest regression. By incorporating features, which consider a long range neighborhood, high level structures are predicted as object *resp.* background, which significantly improves the statistical shape model based segmentation.
- **Appearance Model and an Implicit Shape Model**
  The reference object, represented as a label map, is deformed by a statistical deformation model. Such an implicit model allows to easily combine the modeled shape variation with the appearance model.
- **Incorporate Regions and Object Boundary**
  We propose an image similarity metric for the model fitting, which incorporates the regional object and background probability as well as the object boundary.
- **Surface Fitting to Object Boundary**
  We apply the boundary part of the metric in surface model fitting, where for instance the geodesically damped surface model is used.

1.5 RELATED WORK

Non-rigid image registration has been extensively studied in literature and several attempts have been made to reach a general framework for different smoothness priors. In the following, the literature of this field is reviewed, focused to the work in hand. For a more detailed study of deformable image registration we refer to the survey of Sotiras et al. [91]. Subsequently, we examine the related work to statistical shape models, with the focus to the application of segmentation. A comprehensive overview can be found in Tsechpenakis [95].

REGISTRATION The first attempts to non-rigid and nonparametric image registration go back to the 90's. The deformations have been modeled in terms of linear elasticity, which is described by the Navier-Cauchy partial differential equation (PDE) (*cf.* Evans [31], McOwen [66] and Modersitzki [67] for a thorough study of PDEs). The smoothness prior, better known as the regularization, is implemented by the convolution of the differential operator's Green's function with the force field [19].

Elastic Image Registration

The shortcoming in this approach is its computational demand. In the popular and very efficient *Demons* algorithm (Thirion [94], Pennec et al. [74]) the registration is modeled in terms of a diffusion process with the price of giving up the elastic property of the resulting deformations. However, with an efficient local linearization scheme, the regularization could be implemented by fast Gaussian filtering. In the same manner, Long et al. [60] proposed to implement the regularization by filtering with the exponential kernel, which better approximates the Green's function of the linear elasticity operator. But, the filtering remains expensive, since the convolution could not be efficiently implemented separable, as it is done for the convolution with a Gaussian. Beuthien et al. [14] proposed an efficient separable filter approximation scheme for the filtering with arbitrary stationary kernel functions. But, their approach only works efficiently if the filter kernel is separable. In our work, we use a low-rank approximation, to more accurately approximate separable filters of arbitrary nonseparable ones to still benefit from separable filtering.

Kernel Functions as Smoothness Prior

Further, another type of regularization have become popular, where the resulting deformations are smoothly invertible. This is in particular meaningful in medical applications, where spatial foldings are unnatural and undesirable. First attempts were presented by Christensen and Johnson [22], where in the registration the forward and reverse transformation are jointly estimated as being inverse of one another. To stay in the efficient *Demons*-framework, in [12, 84, 100], exponential maps have been applied, which ensures the deformations to be diffeomorphic. We also integrate this extension in our approach.

Smoothly Invertible Deformations

In contrast to the diffeomorphic Demons approach [100], Beg et al. [13] introduced the LDDMM framework, which is based on time-varying speed vector-fields. This ensures that the resulting deformations are smoothly invertible *i.e.* diffeomorphic. Schmah et al. [83] extended this approach to spatially varying kernels.

HYBRID REGISTRATION The above mentioned methods are based on local iterative optimization schemes. As such, they rely on reasonable initializations and are prone to get *stuck* in local minima. Furthermore, due to *e.g.* artifacts, the image forces not always lead to the desired correspondence. To cope with these problems, hybrid registration algorithms have been proposed, where the image as well as landmarks are considered. For example in [39, 47, 72], the landmarks are treated as additional constraints. The methods require a perfect interpolation of the landmarks resulting in challenging numerical problems during optimization. Other methods do not enforce the landmark constraints strictly,

Incorporation of Landmarks

but add the landmark differences as another cost term to the minimization functional [62, 71, 87, 90]. In our work, we propose to integrate the landmarks into the regularization, as an additional deformation prior using Gaussian process regression. Advantageous is, that the uncertainty of the landmarks can directly be defined independently of the image similarity. For applying this in the *Demons*-like approach nonstationary filtering is required, since the regularization depends on the spatial location. Using tensor decomposition techniques, we present an efficient way to implement the nonstationary filtering. Locally adaptive regularizations have been proposed also by Cahill et al. [20]. With a coupled PDE system, an efficient image-driven adaptive regularization is derived.

RKHS REGISTRATION An alternative approach to registration is the direct search of a solution in a reproducing kernel Hilbert space (RKHS) [86, 87, 92]. Each positive definite kernel function gives rise to a specific regularization property. The flexibility of the RKHS framework comes at the price of a computational demand that exceeds capabilities in medical image applications. However, using low-rank approximations of the corresponding kernel, we present a very general and efficient registration approach, where arbitrary kernels, as for example an empirical covariance function learned from sample data, are applied, combined and conditioned on landmarks. Unlike the approach of Schölkopf et al. [87], where a generic kernel is used as regularization (*e.g.* a Gaussian kernel) and landmark costs are added as a separate term to the optimization, in our work, we derive a hybrid registration by conditioning the prior kernel on the landmarks. We derive a covariance function of the posterior model using Gaussian process regression.

Low-rank Approximation

The use of Gaussian process models for non-rigid registration has been extensively studied in the 90's by Grenander and Miller [38] (and references therein). The novelty of our work is the use of the Nyström approximation [77] to obtain a low-rank approximation of the corresponding Gaussian process. This leads to an efficient numerical method, for any kernel function which is sufficiently smooth. An analytical form of the kernel's basis functions is not required. In particular, this makes it possible to combine kernel functions for shape model fitting and non-rigid registration.

Subsampling Basis Functions

SHAPE AND APPEARANCE MODELS Since generic regularizers often are not restrictive enough, statistical shape model approaches like Active Shape Models [25] and the Morphable Model [16], which are based

on point distributions, have been extensively used in image segmentation [95].

In parametric approaches, a segmentation is sought within the modeled variation of the reference shape. This process is referred to as model fitting. In approaches based on point distributions [18, 25, 58], appearance profiles orthogonal to the object boundary drive the model fitting. One problem common to these approaches is, that they do not model the object and background explicitly, since they only consider information along the object boundary (*cf.* Schönborn et al. [88]). Region based appearance methods [17, 26, 54] model the whole inside of the object using the learned intensity or appearance distribution of the object from registered training samples. However, no background is modeled. In Van Ginneken et al. [98], the background is considered, but only based on few feature landmarks.

Modelling Appearance and Shape Boundary

Cootes et al. [24] and a similar approach by Lindner et al. [59] use random forest regression voting to drive the evolution of the surface model. However, it has only been applied to 2D images with small surface models of between 17 to 65 points.

Atlas-based segmentation methods [33, 34, 73] try to construct a reference shape, having a representative appearance, which is fitted to unseen data. An advantage of atlas-based approaches is that the optimization is straightforward. However, an atlas has to be selected or constructed and it is difficult to cover the background in an atlas.

Cremers et al. [29] propose a linear model of probabilistic training shapes as a prior and derive a convex energy term for segmentation. The linear model is essentially an eigenface model [97] of label maps, which is fitted to a target object probability map. In our work, however, we consider a deformation model on a reference shape (*cf.* Rueckert et al. [80]), which leads to a nonconvex optimization, but allows to generate shapes of a broader class of objects. In Cremers et al. [29], color histograms are learned to predict object probabilities, whereas in Ruggeri et al. [81], conditional random fields are used for capturing local pixel dependencies for classification. To improve the object prediction, we propose to apply random forest regression using advanced features as for example harmonic filters [82], for considering possibly long range neighborhood.

Implicit Shape Model

Lempitsky et al. [56] introduced the use of random forest classification for image segmentation in ultrasound images. They count on the image intensities and the spatial position as features. Geremia et al. [36] applied random forests directly for image segmentation as well. As features they used mean intensities over displaced asymmetric cuboidal

Random Forest Classification

regions. Margeta et al. [65] invented a layered random forest segmentation scheme. They first normalize the image intensities with the help of random forests, perform a rigid image registration of the target and again learn a random forest for the final segmentation. Similar features have been used as in Geremia et al. [36]. All these approaches, however, rely on the random forest appearance model for segmentation and do not combine their model with a shape model to reach global consistency across the segmentation, as it is done in our work.

Geometric approaches [28, 57, 68] are topology independent and are very successful. But, because of a large degree of freedom, the optimization is not straightforward and computationally intensive [78]. These approaches are not in the scope of this work.

ENHANCED MODEL SPACE Statistical shape models are often too rigid as they do not capture the full variability of an object class. Especially, if the number of training samples is small, compared to the shape variability, the generalization of shape models can be poor. Hence, various approaches have been proposed to relax these model restrictions. For instance, Wang and Staib [102] and Albrecht et al. [8] proposed to add the statistical model prior as soft-constraint to the registration functional. Alternative approaches try to artificially enhance the model span. Lebart [55] introduced a local principal component analysis, where correlations between points of the shape are damped with respect to their spatial proximity. In a similar way, Lötjönen et al. [61] created an artificially enhanced training set of shapes. However, both approaches only work with small shapes, since the full covariance matrix has to be explicitly calculated.

Enhancing the Model Span

Based on the low-rank approximation framework, which we present in this work, we show two different possibilities to enhance the model span. First, the statistical model is combined with a generic kernel, which reduces the model bias. The combination is again a kernel, whose basis functions are approximated within our framework. Second, we construct a kernel, where the correlations between points is damped with respect to their geodesic distance on the surface of the object. As a result, the model is less restrictive and the kernel regularizes distances on the shape surface rather than direct Euclidean distances.

Combining Kernel Functions

1.6 OVERVIEW

This thesis is organized as follows. In the following Chapter 2, the elementary minimization formulation for registration is introduced. Two

approaches for minimization are presented. In Section 2.2 the variational approach for registration is introduced. This is followed by the inhomogeneous regularization approach in Section 2.2.3, which is applied to hybrid landmarks and image registration. In Section 2.4, the low-rank Gaussian process approach for registration is presented, which is applied to hybrid registration as well, using a landmark posterior model. In Chapter 3, the statistical shape model is introduced from the Gaussian process perspective, which matches the low-rank Gaussian process formulation of Chapter 2. This is followed by the nonparametric statistical appearance model in Section 3.2 which is used for statistical model fitting for segmentation. In Chapter 4, the presented approaches are evaluated. We conclude our contributions in the last Chapter 5.

Part II

SHAPE MODELING & OBJECT SEGMENTATION

2 REGISTRATION

In this chapter, we introduce a very general objective functional, which will appear throughout this work. It is interpreted as registration problem in this chapter and it will turn out in the following chapter, that it can also be interpreted as statistical deformable model-based image segmentation problem. To flexibly model smoothness prior, we focus on two kernel-based approaches to minimize that functional. The first is based on the variational principle where a semilinear PDE is approximately solved. We additionally present an inhomogeneous regularization scheme, which makes it possible to incorporate landmarks into the regularization. In the second approach, a solution is directly sought in an RKHS. This allows to incorporate landmarks in the regularization as well, however, an exact solution is computationally costly. Finally, we present an approximation scheme, to efficiently find an approximated version of the optimum in the RKHS.

2.1 PROBLEM DESCRIPTION

Segmentation by Registration

In registration, the goal is to express a new target image by a well known reference image, such that knowledge about the reference can be transfered to the target. More concretely, the point-to-point correspondence between the images is sought, such that this knowledge transfer becomes possible. For instance, if in the reference an object is marked by a label, this label can be transfered to the target image and gives rise for a segmentation. Given a target image $I_T : \Omega \subset \mathbb{R}^3 \to \mathbb{R}$, for each point $x \in \Omega$ in a certain domain Ω the assignment of the label $l \in \{0, 1\}$ that corresponds to the object ($l = 1$) or the background ($l = 0$) is sought, which yields a label map $L_T : \Omega \subset \mathbb{R}^3 \to \{0, 1\}$. If a representative reference image $I_R : \Omega \subset \mathbb{R}^3 \to \mathbb{R}$ and its corresponding label map $L_R : \Omega \subset \mathbb{R}^3 \to \{0, 1\}$ is known, the segmentation problem can be formulated in terms of a registration problem. That means, a displacement field $u : \Omega \subset \mathbb{R}^3 \to \mathbb{R}^3$ is sought, such that each point in the warped reference image semantically corresponds to the point in the target image $I_R(x + u(x)) \mathrel{\widehat{=}} I_T(x), \forall x \in \Omega$. Having found a reasonable u, the reference label L_R can be transfered to the target image where $L_R(x+u(x)) \mathrel{\widehat{=}} L_T(x), \forall x \in \Omega$ is finally the segmentation. In Figure 5, a

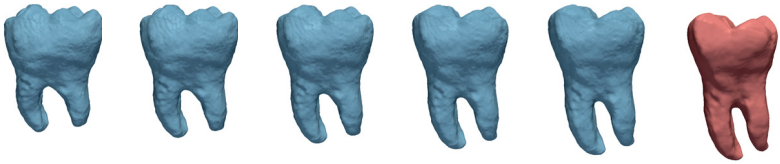
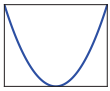


Figure 5: As a result from the registration, a displacement field u is given, which can be used to transform the reference (*blue*) to the target shape (*red*).

reference tooth shape (*blue*) is transformed along u to the target shape (*red*).

LOSS FUNCTIONS For finding an u, which registers I_R and I_T, a matching criterion between the two images is needed. Such a criterion is quantified in terms of a distance function $\mathcal{D}[I_R, I_T, u]$, which measures the dissimilarity between the transformed reference and the target images. This measure can be expressed by a loss function $\mathcal{L} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, which defines the dissimilarity of two point values $I_R(x + u(x))$ and $I_T(x)$. By integrating over the full domain Ω, the corresponding image to image dissimilarity metric becomes

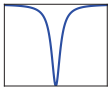
$$\mathcal{D}[I_R, I_T, u] := \int_\Omega \mathcal{L}(I_R(x + u(x)), I_T(x))\mathrm{d}x. \tag{2.1}$$



Squared Loss

A simple and very frequently used loss function is the squared loss

$$\mathcal{L}_s(x, x') := (x - x')^2 \tag{2.2}$$

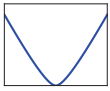


Geman McClure Loss

which strongly penalizes dissimilar values. Hence, it is not robust with respect to *e.g.* artifacts in the image. The Geman McClure loss [35]

$$\mathcal{L}_{GM}(x, x') := \frac{(x - x')^2}{1 + (x - x')^2} \tag{2.3}$$

is a robust loss function, which relaxes the penalty for large differences. Thus, the influence of outliers, as for example artifacts, is reduced. A behavior of something in between these two loss functions results from the Huber loss [45]



Huber Loss

$$\mathcal{L}_h(x, x') := \beta^2 \left(\frac{|x - x'|}{\beta} - \log\left(1 + \frac{|x - x'|}{\beta}\right)\right). \tag{2.4}$$

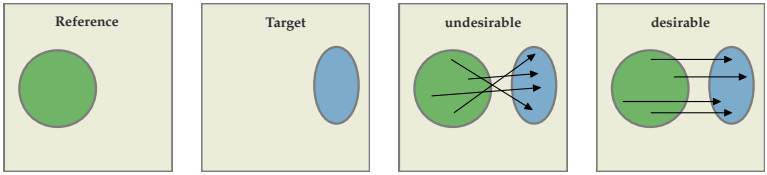


Figure 6: Based on this example, the ambiguity of the correspondence problem can be discussed. Let assume that the background value greatly differs from the values inside the circle or the oval. If we sum up the point value difference of the transformed reference and the target at the four displacement vectors, this leads to the same "cost" in both scenarios *undesirable* and *desirable*. However, in medical image analysis, transformations as the ones in *undesirable* are unnatural and would mean that tissue or bone structures can be folded over. Therefore we prefer transformations as the *desirable* one.

In the experiments in Chapter 4, we apply the squared as well as the Huber loss functions, depending on which kind of images are registered. Further, in combination with the appearance model introduced in this work (*cf.* Section 3.2), we present a more complex similarity measure, where the object and its boundary as well as the background is considered.

2.1.1 *Regularized Minimization Functional*

Minimizing the image distance with respect to u is an ill-posed problem, since no unique minimum exist. For instance, a trivial and undesirable solution would be to find an arbitrary point x' in the target image, where $I_R(x) \approx I_T(x')$ and define $u(x) := x - x'$. Figure 6 illustrates this problem.

Hence, some regularity *resp.* smoothness of the displacement field u is desirable. Instead of simply considering $\mathcal{D}$, in the following regularized minimization functional

$$\hat{u} := \arg\min_{u \in \mathcal{U}} \mathcal{D}[I_R, I_T, u] + \eta \mathcal{R}[u], \tag{2.5}$$

a regularizer $\mathcal{R}$ is incorporated, which additionally measures how well the current solution u fits to the prior assumption about the space of

admissible deformations $\mathcal{U}$. The trade-off between the two terms is controlled by the parameter η.

REGULARIZERS For penalizing *undesirable* deformations as visualized in Figure 6, the following regularizers quantify the smoothness of the sought transformation. The diffusion regularization

$$\mathcal{R}^{\text{diff}}[u] := \frac{1}{2}\sum_{i=1}^{d}\int_{\Omega}\|\nabla u_i(x)\|^2 \mathrm{d}x, \tag{2.6}$$

favors small changes in u *i.e.* smooth displacement fields. The sum in Equation 2.6 goes over all space dimensions d, and u_i indicates the i-th component of u. This regularization is often applied in *Demons*-like registration methods [93, 94]. However, since only the gradients are considered in this transformation model, this can still lead to unnatural expansions *resp.* compressions in the transformation. A soft incompressibility can be ensured by adding the divergence[1] $\nabla\cdot$ of u to the regularization, which is known as the linear elasticity model [7, 19, 67]

$$\mathcal{R}^{\text{elas}}[u] := \mu\frac{1}{2}\sum_{i=1}^{d}\int_{\Omega}\|\nabla u_i(x)\|^2 \mathrm{d}x + \nu\frac{1}{2}\int_{\Omega}(\nabla\cdot u)^2 \mathrm{d}x. \tag{2.7}$$

μ and ν define the stiffness and the compressibility respectively. In registration, the linear elasticity model is often desirable, since it softly prevents folds in the deformations and penalizes volume changes. Later in this work, we approximate this model by using an exponential filter kernel.

2.1.2 *Hybrid Image Registration*

Incorporate Landmark Displacements

The image distance $\mathcal{D}$ does not necessarily reflect the desired image similarity. For instance, high-level features not present in the image or artifacts, may lead to misleading correspondences. Hence, it is important to have the possibility to influence the registration result *e.g.* by providing landmarks. In addition to the images I_R and I_T therefore, lists $X_R = \{x_i^R\}_{i=1}^n$ and $X_T = \{x_i^T\}_{i=1}^n$ of n corresponding points for each

1 The divergence of u corresponds to the linearization of $det[D(x + u(x))] = 1 + \nabla\cdot u +$ (nonlinear term), which is the degree of compression (it is 1 for no compression).

image are considered. The displacements induced by these landmarks are given by

$$\begin{aligned} X =& \{(x_1^T, x_1^R - x_1^T), \ldots, (x_n^T, x_n^R - x_n^T)\} \\ :=& \{(x_1, y_1), \ldots, (x_n, y_n)\}. \end{aligned}$$

In various methods [32, 39, 47, 72], the landmarks are treated as hard constraints to the registration functional. The resulting problem is to find a displacement field that minimizes Equation 2.5, but subject to additional landmark constraints:

$$\arg\min_{u \in \mathcal{U}} \mathcal{D}[I_R, I_T, u] + \eta \mathcal{R}[u] \quad s.t. \quad \mathcal{D}^{\text{lm}}[X_R, X_T, u] \leq \varepsilon, \tag{2.8}$$

where $\mathcal{D}^{\text{lm}}$ measures the landmark distance. If $\varepsilon = 0$, a perfect interpolation of the landmarks is required, which results in numerical problems during optimization. Other methods [62, 71, 87, 89], do not strictly enforce the landmark constraints, but add the landmarks as an additional cost to Equation 2.5:

$$\arg\min_{u \in \mathcal{U}} \mathcal{D}[I_R, I_T, u] + \tau \mathcal{D}^{\text{lm}}[X_R, X_T, u] + \eta \mathcal{R}[u], \tag{2.9}$$

where τ is a weighting parameter. This approach has the advantage that the optimization problem is straight-forward to integrate into existing registration schemes. Moreover, we argue that it is more natural to treat the landmarks as soft constraints, since they are usually approximately known only. Later in this work, in Section 2.2.4, we will consider the landmarks in the regularization term of Equation 2.5.

2.1.3 *Shape Representations*

The very general formulation introduced above allows for various registration strategies. Beside CT image to image registration, where *e.g.* the squared or the Geman McClure loss function can be used, the framework allows also for surface registration, where an object is represented as a label map $L : \mathbb{R}^d \to \{0, 1\}$. The object is indicated as 1 and the background as 0. The boundary of the object is defined as the surface $\Gamma := \{x | x \in |\nabla L(x)| > 0\}$. For the registration of the reference and a target surface, Γ_R and Γ_T can be represented as distance maps obtained by the signed Euclidean distance transform which is defined as

$$I^D(x) = \begin{cases} dist(x, \Gamma) & \text{if } x \in \text{outside}(\Gamma) \\ 0 & \text{if } x \in \Gamma \\ -dist(x, \Gamma) & \text{if } x \in \text{inside}(\Gamma). \end{cases} \tag{2.10}$$

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

Figure 7: Example of a: (a) CT image, (b) label map, (c) distance image and (d) tooth probability map.

Registering the two distance maps I_R^D and I_T^D by minimizing Equation 2.5 corresponds to a closest point surface registration. For building the statistical shape model, later in this work, we used this representation.

Probabilistic Shape Functions

One further type of representation finds application in this work. As a generalization of the label map, we define the probability map $P : \mathbb{R}^d \to [0, 1]$, which indicates the probability of a point in the image being part of the object. For image segmentation we use target probability maps. In Figure 7, examples of the different representations are shown.

In the following, two different strategies are discussed on how the registration functional Equation 2.5 can be minimized. In Section 2.2 and Section 2.2.3, we introduce the variational approach, which leads to solving a system of semilinear PDEs and in Section 2.4, we directly minimize Equation 2.5 in a reproducing kernel Hilbert spaces.

2.2 VARIATIONAL REGISTRATION

Minimizing the regularized functional in Equation 2.5 is a nonlinear optimization problem, since the nonrigid transformation to the images is a nonlinear operation. To find a local optimum, the variational approach turned out to be very useful in literature [14, 19, 93, 103]. We now introduce the essentials of the approach and refer to Evans [31] and Modersitzki [67] for more details.

A displacement field $u : \mathbb{R}^d \to \mathbb{R}^d$ that registers the two images I_R and I_T is sought as the minimum of the joint functional

$$\mathcal{J}[u] = \mathcal{D}[I_R, I_T, u] + \mathcal{R}[u]. \tag{2.11}$$

Using methods from calculus of variations, the functional is differentiated with respect to transformation u. If the space of all admissible transformations is denoted as Φ, the functional derivative of $\mathcal{J}$ becomes the sum of the derivatives $d\mathcal{D} = f : \Phi \to \mathbb{R}^d$ and $d\mathcal{R} = A : \Phi \to \Phi$. A is usually restricted to be a linear self-adjoint differential operator which leads to the following system of semilinear[2] partial differential equations

Functional Derivative

$$Au = f(u), \tag{2.12}$$

whose solution minimizes the functional (2.11). Equation 2.12 is called the Euler-Lagrange equations of Equation 2.11, where A measures the regularity of u and f is referred to as force field. (see Section A.6 for more details about functional derivatives). For the squared loss for instance the force field is give as

$$f(u)(x) = 2(I_R(x + u(x)) - I_T(x))\nabla(I_R(x + u(x))). \tag{2.13}$$

Because it contains the warped reference image $I_R(x + u(x))$, the term f depends nonlinearly on u. This makes an analytical solution of the system in Equation 2.12 intractable.

Image Forces

An efficient method to approximately solve Equation 2.12, which has well established in 3D medical image registration [14, 19, 93, 94], is by find a solution to the problem Equation 2.12 using the "gradient flow". It can be interpreted as a continuous gradient descent for minimizing Equation 2.11. A time variable is introduced which leads to a time-dependent partial differential equation

Gradient Flow

$$\begin{aligned} \frac{\partial}{\partial t}u - Au &= f(u) \\ u_t &= 0 \quad \text{for} \quad t = 0, \end{aligned} \tag{2.14}$$

whose stationary solution $u^*(x)$ solves Equation 2.12, since the time derivative $\frac{\partial}{\partial t}u^*$ vanishes *cf.* the work of Evans [31] and Nolen [69]. By Duhamel's principle, the solution to the nonhomogeneous initial-value problem Equation 2.14 is

$$u_t = \int_0^t T_s f(u_s)\mathrm{d}s, \tag{2.15}$$

where $(T_s f)(x) = \int_\Omega k_s(x-y)f(y)\mathrm{d}y$ is the solution to the homogeneous problem

$$\frac{\partial}{\partial t}u - Au = 0, \quad u_s = f(u_s). \tag{2.16}$$

2 Equation 2.12 is called to be semilinear because A is a linear operator but f depends on u.

Fixed-Point Iteration

$k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is called the fundamental solution of $\frac{\partial}{\partial t} - A$ and is a positive definite kernel function if A is linear, self-adjoint and positive definite (*cf.* Steinke and Schölkopf [92] and Ascher et al. [11]).

Since f still depends nonlinearly on u, Equation 2.15 is iteratively solved using Picard's iteration process [21]

$$u_t^{i+1} = \int_0^t T_s f(u_s^i) \mathrm{d}s, \quad u_t^0 = 0. \tag{2.17}$$

For each time step the forward Euler method yields the final fixed-point iteration scheme

$$u^{i+1} = k * (u^i + f(u^i)). \tag{2.18}$$

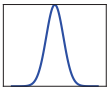
In each iteration, the previous solution u^i is known and there is no nonlinear dependence on u^{i+1}. In fact, Equation 2.15 is linearized by evaluating f at u^i and not at u^{i+1}. The iterative scheme Equation 2.18 amounts to solving the time-dependent partial differential equations Equation 2.14 and can be seen as a discrete gradient descent of the initial functional Equation 2.11, in which the "gradient" is efficiency handled by the convolution operator $k*$.

2.2.1 *Linear Operators*

In the classical example of the Demons algorithm [94], the diffusion regularization term $\mathcal{R}^{\text{diff}}$ (Equation 2.6) finds application. Its derivative is given as

$$A^{\text{diff}} = -\Delta. \tag{2.19}$$

Using A^{diff} in Equation 2.14, the PDE is called the reaction diffusion equation[3]. The fundamental solution of the homogeneous time-dependent PDE is given as the Gaussian kernel



Gaussian Kernel

$$k_G(x, y) = \frac{1}{(\sqrt{2\pi}\sigma)^d} \cdot \exp\left(\frac{-\|x - y\|^2}{2\sigma^2}\right). \tag{2.20}$$

Since the Gaussian kernel is separable, the convolution in Equation 2.18 can be very efficiently implemented by separate 1D convolutions.

The derivative of the linear elasticity regularization operator $\mathcal{R}^{\text{elas}}$ (Equation 2.7) is given as

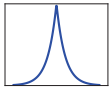
$$A^{\text{elas}} = -\mu\nabla^2 - \nu\nabla\nabla\cdot. \tag{2.21}$$

3 The PDE is called "reaction" diffusion equation because the right-hand side f depends on u, which makes the system semilinear.

Based on the eigen-function basis of A^{elas} (derived in [23]), Bro-Nielsen and Gramkow [19] approximated a convolution kernel of this operator. Since, the computation of this kernel is not cheap, Long et al. [60] propose to efficiently approximate it by the exponential kernel

$$k_e(x,y) = \frac{1}{C_d} \cdot \exp\left(\frac{-\|x-y\|}{\alpha}\right), \tag{2.22}$$

where $C_1 = 2\alpha, C_2 = 2\pi\alpha^2, C_3 = 8\pi\alpha^3$ and d is the space dimension. However, since both kernels are not separable, the convolution becomes computationally demanding and is not directly applicable in usual medical image registration. Later in this chapter, we will present a convolution scheme based on separable filter approximation, which enables an efficient convolution with the exponential kernel.



Exponential Kernel

CHOOSING KERNEL FUNCTIONS In variational image registration, the regularization outcome fundamentally depends on the choice of the operator A, since it defines the regularity the transformation has to satisfy. Unfortunately, the fundamental solution of only few regularizers are analytically known, whilst it is hard to derive k given a particular regularization operator. In principle, however, it is sufficient to find the particular kernel function, which leads to the favored regularization property. Therefore, the specific regularization operator does not have to be explicitly known. Later in this work, we will construct a landmark kernel without actually knowing the corresponding regularization operator.

Knowing the Regularization Property

2.2.2 *Neighborhood Preservation*

Common simple regularizers, as the ones discussed above, penalize unsmooth transformations. But, using the iteration scheme of Equation 2.18, foldings (*cf.* Figure 6) cannot be avoided. In medical application however, foldings are unnatural and undesirable, since that would mean, that tissue and bone structures can be folded over. If we restrict the resulting transformation to be smoothly invertible *i.e.* diffeomorphic, guarantees that there are no foldings in the transformation. This is why in medical applications invertible transformations are preferred. Various authors [10, 13, 22, 84, 100] have therefore addressed an additional restriction of the mappings $u^i \in \text{Diff}(\Omega)$ to be diffeomorphic. Following the diffeomorphic Demons approach [100], the restriction of an optimal displacement field to be diffeomorphic can be achieved by mapping the current transformation u^i back onto the Lie group of diffeomorphisms. This ensures

Diffeomorphic registration

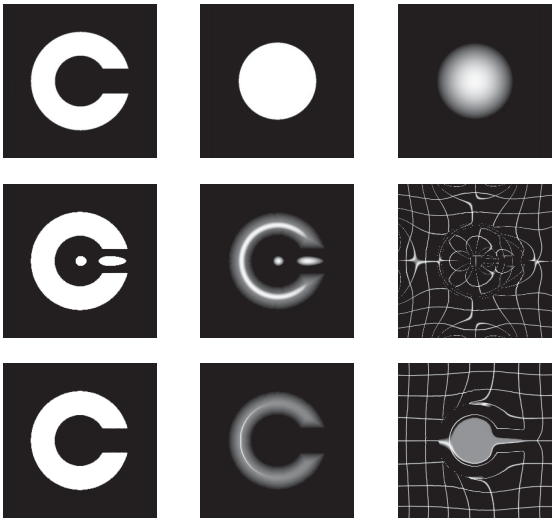


Figure 8: In this toy example, a circle is registered onto a C. With this example, the difference between "non-diffeomorphic" and diffeomorphic registration can be well illustrated. In the first row, the target (C) and the references (circle and shaded circle) are shown. In the second row, the standard Demons registration method [94] was used, while in the third row the diffeomorphic Demons [100] was applied for registration. From left to right: the warped reference circle, the warped shaded circle and the corresponding warped grid is depicted. In the standard Demons example, foldings lead to strong artifacts, while in the diffeomorphic Demons example, the circle is "naturally" deformed.

the transformations to be smoothly invertible and thus neighborhood preserving. A possible way to do that, which only marginally changes the optimization scheme in Equation 2.18, is by calculating the group exponential map exp of the Lie group of diffeomorphisms for u^i after evaluating the field update

$$u^{i+1} = k * (u^i + \exp(f(u^i))). \tag{2.23}$$

The exponential mapping can be efficiently approximated by a scaling and squaring algorithm [10, 100]. In Figure 8, the impact of the diffeomorphic constraint is visualized using a toy registration example.

2.2.3 *Inhomogeneous Regularization*

The regularization introduced above is homogeneous *i.e.* independent of the spatial location. This bases on the assumption that, in the full domain of the image, the same smoothness requirement is true. In medical image registration, where tissues and bone structures with different deformable properties are present in the image, the homogeneity assumption is not true anymore. To include such image driven cues, Weickert et al. [103] introduced anisotropic diffusion which adapts to image gradients. Cahill et al. [20] presented an image-driven locally adaptive regularization, where the degree of regularity is weighted inversely proportional to the local gradient magnitude of the image.

Locally Adaptive Regularization

Especially, if landmarks are given, *e.g.* in regions of artifacts, the homogeneity assumption is also not true anymore. In the vicinity of the landmarks, the transformation is known, thus, a higher smoothness is required at this location during registration. In contrast to image-driven regularization weighting, in this section, we present a kernel, which incorporates given landmarks and thus changes the regularization depending the image location. This property is called nonstationarity (*i.e.* not translation-invariant). To apply this kernel in the present approach, we generalize the variational image registration framework to kernels functions, which possibly are nonstationary.

FREDHOLM INTEGRAL EQUATIONS For the generalization of Equation 2.18 to nonstationary kernels, the optimization scheme has to be rewritten. We explicitly write the convolution integral, but for kernels which are not stationary

$$u^{i+1} = \int_{\Omega} k(\cdot, s)(u^i + f(u^i))\mathrm{d}s. \tag{2.24}$$

Similar, with u restricted to be diffeomorphic, Equation 2.23 becomes

Inhomogeneous Fredholm equation

$$u^{i+1} = \int_{\Omega} k(\cdot, s)(u^i + \exp(f(u^i)))\mathrm{d}s. \tag{2.25}$$

Note that now, k depends on both arguments and not only on their difference $k(x - y) \neq k(x, y)$. The integral equation Equation 2.24 is known as the inhomogeneous Fredholm equation. For more extensive work on this type of equations we refer to Pipkin [76] and Kress et al. [52].

2.2.4 *Incorporating Landmarks*

In literature, landmarks are often incorporated using an additional cost term in the registration functional [62, 87]. Thus, the landmarks and the regularization are treated independently, even though landmarks clearly provide a-priori information about the deformations. Further, parameter tuning becomes difficult, since the weighting of the landmark cost and the regularization term are mutually dependent. In the following, we construct a kernel which is aware of the prior landmark displacements and thus incorporate the landmarks into the regularization. The only additional parameter we will introduce is the uncertainty σ of the landmark positions, which is usually approximately known.

Gaussian Processes

For constructing a kernel, which incorporates prior displacements (landmarks), we use the concept of Gaussian processes, which is strongly related to kernel functions. A Gaussian process defines a probability distribution over a function space. It is a collection of random variables $u(x), x \in \Omega$, where Ω is an index set, with the property that for any finite number of observations, $x_1, \dots, x_n \in \Omega$, the values $u(x_1), \dots, u(x_n)$ are jointly normally distributed. Without loss of generality, scalar-valued kernels are considered where each space dimension is treated independently. For the generalization to matrix-valued kernels see Section 2.4.1.

A Gaussian process is completely defined by a mean function $\mu : \Omega \to \mathbb{R}^d$ and a covariance function $k : \Omega \times \Omega \to \mathbb{R}$. By specifying a covariance function k, it is defined which functions are likely under the given process. The covariance function specifies for each pair of points x, y their covariance and is often referred to as kernel. In fact, covariance functions are exactly the same mathematical object. Throughout this work, we will use the term kernel function.

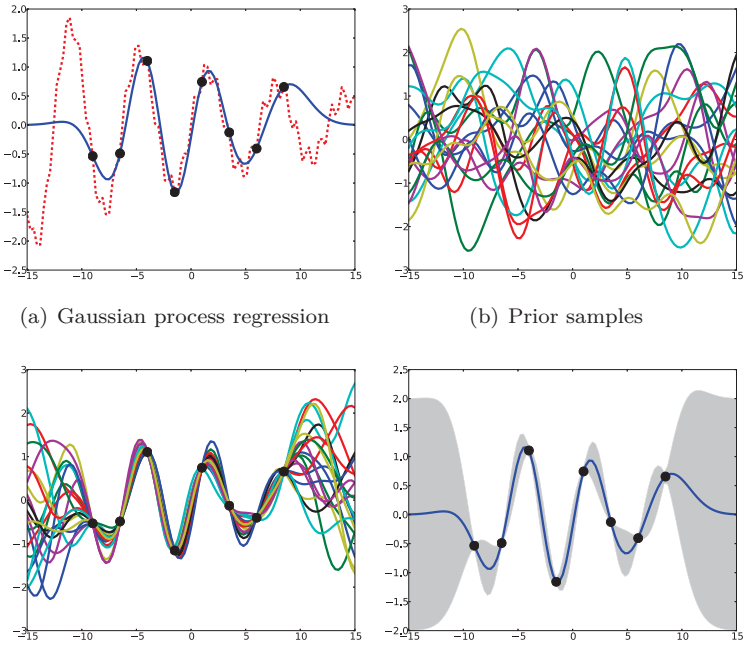
Regression and Uncertainty

Assume an i.i.d. sample $X = \{(x_1, y_1), \dots, (x_n, y_n)\} \subset \Omega \times \mathbb{R}^d$ is given and let $u \sim \mathcal{GP}(\mu, k)$ be a Gaussian process with some mean $\mu : \Omega \to \mathbb{R}^d$ and a kernel $k : \Omega \times \Omega \to \mathbb{R}$. Let us infer the distribution $p(u|X)$. Additionally, let assume that $y \sim \mathcal{N}(u(x), \sigma^2 \mathbf{I}_d)$. That means, instead of observing the actual values $u(x)$, noisy instances y thereof are observed. Under this assumption, the posterior distribution $p(u|X) \propto p(u)p(X|u)$ is known in closed form and is again a Gaussian process $\mathcal{GP}(\mu_X, k_X)$

$$\mu_X(x) = \mu(x) + K_X(x)^T (K_{XX} + \sigma^2 \mathbf{I}_n)^{-1} Y \tag{2.26}$$

$$k_X(x, x') = k(x, x') - K_X(x)^T (K_{XX} + \sigma^2 \mathbf{I}_n)^{-1} K_X(x'). \tag{2.27}$$

Here, $K_X(x) = (k(x, x_i))_{i=1}^n \in \mathbb{R}^n$, $K_{XX} \in \mathbb{R}^{n \times n}$ is the kernel matrix with entries $K_{i,j} = k(x_i, x_j)$ and $Y = (y_1 - \mu(x_1), \dots, y_n - \mu(x_n))^T \in \mathbb{R}^n$ are the mean free landmark displacements and σ^2 models the uncertainty



(a) Gaussian process regression

(b) Prior samples

(c) Posterior samples

(d) Confidence interval

Figure 9: Gaussian process regression example in 1D. The observed function in (a) colored in *red* is $y(x) = \sin(x)\exp(-0.05x) + 0.2\sin(6x) + \epsilon$. The posterior mean μ_X given the landmark points (black dots) is plotted in *blue*. In (b), samples are plotted from $\mathcal{GP}(0, k_G)$, where k_G is the Gaussian kernel with $\sigma = 2$. In (c), samples are plotted from the posterior process $\mathcal{GP}(\mu_X, k_X)$ and in (d), the confidence interval of two standard deviations is depicted as *gray* shaded area.

about matching accuracy of the landmarks (see also *e.g.* Rasmussen [77], Chapter 2.2).

In Figure 9, an example Gaussian process in 1D is shown. Since the full posterior distribution is known, a confidence interval can be given and samples can be drawn from $p(u)$ *resp.* $p(u|X)$.

GAUSSIAN PROCESSES FOR REGISTRATION Let the prior knowledge about the deformations u in Equation 2.11 be modeled using a Gaussian process $u \sim \mathcal{GP}(\mu, k)$, which is defined by a mean function $\mu : \Omega \to \mathbb{R}^d$ and a kernel function $k : \Omega \times \Omega \to \mathbb{R}$. We adjust Equation 2.11

$$\mathcal{J}_{GP}[u] := \mathcal{D}_{GP}[I_R, I_T, \mu, u] + \mathcal{R}_{GP}[k, u], \tag{2.28}$$

Hybrid registration

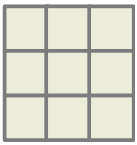
where the target image is warped by $I_T(x + \mu(x) + u(x))$. The strength of this generalized interpretation is that now the hybrid registration problem can be formulated by conditioning the Gaussian process on the given landmark displacements. Hence, the functional which is minimized becomes

$$\mathcal{J}_{GP_X}[u] := \mathcal{D}_{GP_X}[I_R, I_T, \mu_X, u] + \mathcal{R}_{GP_X}[k_X, u], \tag{2.29}$$

using μ_X as landmark based mean transformation and the kernel function k_X for regularization. In this context, we name k_X landmark kernel.

Since the landmark kernel depends on the landmark displacements, it is not stationary and the inhomogeneous optimization scheme Equation 2.25 is required.

2.3 DISCRETE FILTERING APPROACH

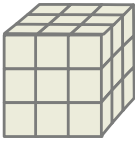


2D Filter

So far, we have derived a continuous formulation to minimize the registration functional. For stationary regularizers, an iterative convolution scheme Equation 2.18 has been introduced, which is of wide use in medical image registration. We generalized the formulation to nonstationary regularizers (Equation 2.25) to enable kernels which depend on the image location. In this section, we discretize the two formulations in order to implement this registration approach.

We start by writing the spatially discretized version of Equation 2.18 where the kernel k becomes a filter mask H_0:

$$u^{i+1}(x) = H_0 * (u^i + \exp(f(u^i)))_x. \tag{2.30}$$



3D Filter

H_0 is the discretized kernel function k at location 0 with elements $H_{0_{ijq}} = k(0, (i, j, q)^T)$ and i, j, q cover the neighborhood around 0. The subscript x of the second term indicates the equally large discrete neighborhood around the point x. In the two dimensional case, the filter mask H_0 is a matrix, while in the 3D case H_0 becomes a third order tensor.

If the Fredholm integral Equation 2.25 is spatially discretized

$$u^{i+1}(x) = H_x * (u^i + \exp(f(u^i)))_x, \tag{2.31}$$

where H_x is the discretization of k at location x, *i.e.* $H_{x_{ijq}} = k(x, (i, j, q)^T)$, H depends on x, which makes the problem nonstationary. In general, all the local filter kernels have to be calculated in order to approximate the Fredolm integral.

Spatial filtering, especially in 3D is computationally costly. Moreover, nonstationary filtering is computationally even worse. In the following sections, we present two approaches to accelerate on the one hand the stationary and on the other hand the nonstationary filtering.

2.3.1 *Separable Filter Decomposition*

The iteration scheme of Equation 2.30 can be accelerated greatly, if the convolution can be performed separately in each space dimension by successive 1D convolutions. In the two dimensional case the separable convolution is given as

$$\begin{aligned} H_0^{2\mathrm{D}} * f &= \left(H_0^{1\mathrm{D}} \otimes H_0^{1\mathrm{D}}\right) * f \\ &= H_0^{1\mathrm{D}} * \left(H_0^{1\mathrm{D}} * f\right), \end{aligned} \tag{2.32}$$

where $H_0^{2\mathrm{D}}$ denotes the 2D and $H_0^{1\mathrm{D}}$ the 1D filter kernel and $\otimes$ the outer product. In the separable convolution, the associativity (2.32) of the convolution operation is exploited. For each output pixel, the computational complexity is reduced from m^2 to $2m$, where m is the width of the filter. However, this is only possible if the filter kernel is separable *i.e.* $H_0^{2\mathrm{D}} = H_0^{1\mathrm{D}} \otimes H_0^{1\mathrm{D}}$. The Gaussian kernel has this nice property of separability. Therefore, without any further effort, the convolution with this kernel can be performed separately.

Tensor Decomposition

To still benefit from this performance gain for nonseparable kernels, like the exponential kernel [60], separable filters of them have to be approximated. In 2D, this can be achieved by low-rank approximation using standard singular value decomposition. However in 3D, this leads to mathematical challenges that go beyond standard linear algebra, since a filter kernel in 3D is a third order tensor. In contrast to 2D matrices, it is an NP-hard problem to determine the rank of a specific given higher order tensor (see Kolda and Bader [50]). Hence, the rank R becomes a parameter which has to be estimated. Nevertheless, we are able to compute the approximation using the Candecomp/Parafac (CP) decomposition model [50]. This gives us separable 1D approximations of the discrete filter kernel H_0. In Figure 10, the decomposition model is visu-

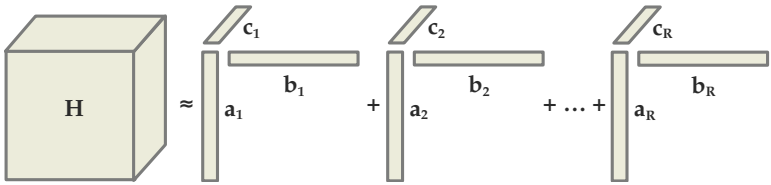


Figure 10: Candecomp/Parafac tensor decomposition model.

ally illustrated. The decomposition can be formulated as a minimization problem

$$\arg\min_{a_r,b_r,c_r} \|H_0 - \tilde{H}_0\| \quad \text{with} \quad \tilde{H}_0 = \sum_{r=1}^{R} a_r \otimes b_r \otimes c_r, \tag{2.33}$$

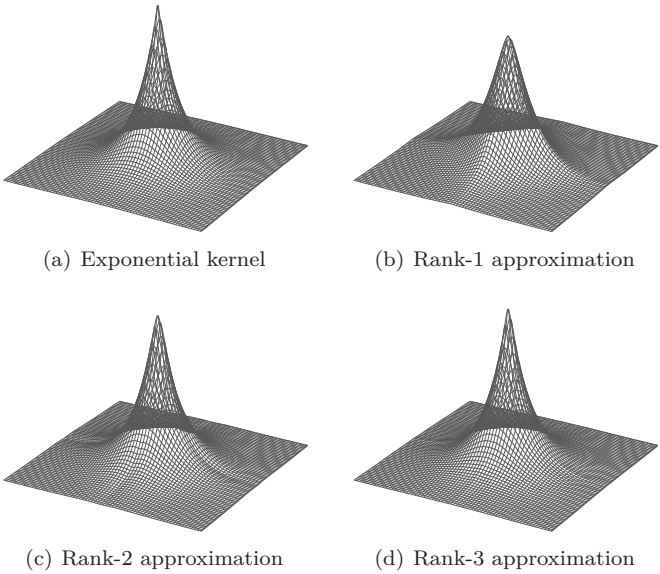
where the operation $\otimes$ denotes the three-way outer product $\tilde{H}_{0ijq} = \sum_r^R a_{ir} b_{jr} c_{qr}$. Standard algorithms to optimize Equation 2.33 are based on the alternating least squares (ALS) method [41], which is explained in more detail in the appendix (Section B.1). The parameter R is estimated by testing the approximation performance for different ranks.

Separate Convolution

Once the decomposition is performed, the distributivity (2.34) and again the associativity (2.35) property of the convolution is exploited to perform the convolution separately with a_r, b_r and c_r

$$\begin{aligned} H_0 * f \approx \tilde{H}_0 * f &= \left[\sum_r^R a_r \otimes b_r \otimes c_r\right] * f \\ &= \sum_r^R \left[a_r \otimes b_r \otimes c_r * f\right] \qquad (2.34) \\ &= \sum_r^R a_r * (b_r * (c_r * f)) \,. \qquad (2.35) \end{aligned}$$

For cubic filter kernels, having a filter width m, the computational cost for each output pixel reduces significantly from m^3 to $3Rm$. In Figure 11, a kernel reconstruction of the 2D exponential kernel with different approximation ranks is depicted. The higher rank R of the approximation is, the more accurate the kernel can be reconstructed.



(a) Exponential kernel (b) Rank-1 approximation

(c) Rank-2 approximation (d) Rank-3 approximation

Figure 11: In this figure, the 2D exponential kernel and its different approximations are depicted. The higher the rank of the approximation the better, the original kernel is approximated.

2.3.2 *Efficient Nonstationary Filtering*

For spatially varying filter kernels, where the nonstationary filtering scheme Equation 2.31 is used, separable filtering is not possible since the associativity of the convolution operation no longer holds. Hence, for each output pixel, the filter mask has to be computed and the convolution at this location has to be performed, which is computationally costly applying 3D medical images. However, in the particular case where the hybrid registration functional in Equation 2.29 is minimized, following properties of the landmark kernel k_X (Equation 2.27) can be exploited to reach an optimization scheme which is computationally feasible.

Landmark Kernel Properties

The landmark kernel k_X consists out of the kernel k subtracted by a landmark dependent term. The difference between k and the full landmark kernel k_X becomes negligible if

$$\forall x_i \in X \; k(x, x_i) < \xi \tag{2.36}$$

i.e. if x is not in the neighborhood of any landmark. This property is used to approximate the integral of Equation 2.31 by only considering k, the first part of the landmark kernel k_X, if the value of its second part goes to zero. We perform the approximation in two steps:

1. At first, the whole image is filtered separately using the stationary part k.
2. Subsequently, the nonseparable and nonstationary filtering with the full kernel k_X is performed, but only for pixels where (2.36) is not fulfilled.

The second part is the most expensive step, because for each point in the vicinity of the landmarks the discretization H_x has to be calculated. This means a cubically increasing amount of kernel evaluations, which covers the neighborhood of all points having landmark support, and this in each iteration of Equation 2.31. To reduce the computational demands, we propose the following caching scheme.

LOCAL FILTER CACHING Since the landmark kernel is nonstationary, but still time-invariant, it is reasonable to keep the computed filter kernels in memory to save computational time for the following iterations. However, the amount of memory to cache all the filter kernels grows rapidly depending on the filter width and the number of landmarks. Therefore, we propose to compress these local filter kernels by again taking advantage of tensor decomposition, before they are cached in the memory.

Compressing Local Filter Kernels

As we have seen in Section 2.3.1, the CP decomposition is obtained using the ALS method, which is costly due to its iterative nature. Because H_0 has to be decomposed only once, it is still well suited to approximate the separability of the stationary filter. But the ALS method is too slow to decompose all the local filters H_x.

Compared to the CP decomposition, the Tucker decomposition [96] is significantly faster. It is an alternative model to decompose a tensor (see Figure 12). Similar to the CP model, the tensor is decomposed into triplets of vectors, but they are weighted by a full so called "core" tensor

$$\tilde{H}_x = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} g_{pqr} a_p \otimes b_q \otimes c_r, \tag{2.37}$$

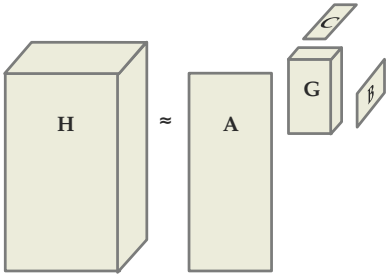


Figure 12: Tucker tensor decomposition model.

where g_{pqr} are the elements of the core tensor G and P, Q, R are the ranks for each space dimension. Using the Higher-order SVD algorithm of De Lathauwer et al. [30]

$$\underset{G, a_p, b_q, c_r}{\arg\min} \|H_x - \tilde{H}_x\| \tag{2.38}$$

can be very efficiently minimized (see Section B.2 for more details).

Compared to the CP model, the Tucker decomposition is a less restricted model where the core G can be dense, while in the CP model, the core is a super diagonal tensor[4] with ones on the diagonal. Although it cannot be used for separable filter approximation due to the weighting with the dense G, the memory savings are similar to the CP model. Setting $P = Q = R$ and having a filter length m, the memory consumption reduces from m^3 to $R^3 + 3Rm$ per voxel in the support of the landmarks. Similar to the parameter R in the CP decomposition, P, Q, R have to be estimated by testing the approximation performance.

Saving Memory

2.3.3 *Multi-Resolution*

The presented method is mainly based on the local iterative minimization scheme Equation 2.31. As such, it relies on a reasonable initialization and is prone to getting "stuck" in local minima. This is a well-known problem in image registration and is typically tackled by adopting a multi-resolution strategy. By first solving the registration problem on low-resolution approximations of the images, a good initialization for

4 A super diagonal tensor is the generalization of a diagonal matrix to higher order tensors, where the entries outside the main diagonal are zero.

the high resolution images is computed. As an added benefit, the low-resolution registrations are computationally less intensive.

An image pyramid is built

$$I^{l-1} = \text{downsample}(G * I^l), l = 1, \ldots, L \tag{2.39}$$

for the reference and target image I_R, I_T. L is the number of levels in the pyramid. I^L is the original high resolution image and G is a Gaussian smoothing kernel. The iteration scheme in Equation 2.31 is first computed at resolution level 0. Each higher level is initialized by the preceding result, the displacement field u, up-sampled to the next higher resolution. This approach is widely used in image analysis but in combination with the posterior mean function μ_X, more effort is needed. The problem is, that the support in voxels of the kernel function k implicitly increases towards the lower resolution levels. At the same time, the mean function μ_X only depends on the landmark points and is independent of the spacing and sampling rate of the images. Therefore, the kernel parameters for calculating the mean μ_X and for regularizing the displacement fields u^l are different in lower resolution levels. In order to harmonize that, we propose to use a kernel that simultaneously represents multiple scales, which can be achieved by a linear combination of different single scaled kernels (*cf.* Opfer [70]). The landmark based mean transformation μ_X is computed using the multiscale kernel

Harmonize Mean on Multiple Scales

$$\tilde{k}(x, y) = \sum_{l=0}^{L} \lambda_l k^l(x, y), \tag{2.40}$$

where the λ_l are positive weights and k^l correspond to k with adjusted kernel parameters per scale level l. The parameters *e.g.* for the Gaussian kernel becomes $\sigma_G^l = 2^{L-l}\sigma_G$. The weights λ_l are equivalent to the scales of the Gaussian kernels.

2.4 RKHS REGISTRATION

In this section, we will introduce a different view onto the registration problem where the space of admissible deformations is restricted to functions in a reproducing kernel Hilbert space (RKHS) [87]. Similar to the variational approach, we restrict the regularization operator A to operators, where a corresponding kernel function k exists. Contrary to the variational approach discussed in the previous section, where a stationary solution to a system of semilinear partial differential equations is sought, the objective functional Equation 2.5 is directly minimized in

the RKHS. The key difference is, that in the variational approach, the fundamental solution to a time-dependent PDE admits a smooth solution to the Euler-Lagrange equations, while in the reproducing kernel Hilbert space approach, the kernel function k of A^*A leads to a solution of the initial registration functional. By expressing the deformations u in the basis of the RKHS induced by this k, the solution is regular with respect to A.

The variational approach for image registration became popular in the 90' [19, 94], while the RKHS framework is still rarely used for image registration [64, 87]. In this section, we intend to emphasize the key difference between the approaches and we show how the RKHS framework can be applied for efficient image registration. Subsequently, we will put the RKHS formulation in context of Gaussian processes to flexibly model new regularization properties.

2.4.1 *Minimizing in the RKHS*

In the following, we summarize the essentials of the variational approach to then introduce the RKHS approach. Recall the objective functional for registration

$$\mathcal{J}[u] = D[I_R, I_T, u] + \eta \mathcal{R}[u]. \tag{2.41}$$

Here, we restrict the regularization term to have the form

$$\mathcal{R}[u] = \frac{1}{2}\|Au\|^2 \tag{2.42}$$

with A as a positive definite linear operator which has an adjoint operator A^*. In Section 2.2, the minimization strategy is to derive the functional derivative of Equation 2.41 with respect to u, and thus to solve the Euler-Lagrange equations

$$A^*Au = f(u). \tag{2.43}$$

The solution to Equation 2.43 yields a minimizer for Equation 2.41. A solution of Equation 2.43 is also the steady state of the semilinear evolution equation

$$\frac{\partial}{\partial t}u - A^*Au = f(u), \quad u_t = 0 \quad \text{for} \quad t = 0. \tag{2.44}$$

Since f depends nonlinearly on u, the system is linearized, which yields a fixed-point iteration scheme

$$u^{i+1} = k * (u^i + f(u^i)), \quad u_0 = 0, \tag{2.45}$$

which converges to a minimizer of Equation 2.41. If A spatially varies, the convolution becomes an integral operator $(T_k f)(x) = \int k(x,y)f(y)\mathrm{d}y$. Further, as long as the linear operator in Equation 2.44 has the form A^*A, k is a positive definite kernel function [31, 92].

In Section 2.2, we argued that, depending on k, different time-dependent operator equations are solved without necessarily knowing their exact form. Note that in each iteration step the convolution with the kernel k ensures that u_{t+1} satisfies the regularity induced by the linear operator whose kernel function is exactly k. In case of the Gaussian kernel the linear operator has the form $\frac{\partial}{\partial t} - \Delta$ and $u_{t+1} \in C^\infty$.

In the RKHS approach, a space of functions is defined

$$\mathcal{H}_k = \Big\{ u | u(x) = \sum_{i=1}^{\infty} \alpha_i k(x_i, x), \qquad x_i \in \Omega, \quad \alpha_i \in \mathbb{R}, \quad \|u\|_k < \infty \Big\}. \tag{2.46}$$

Further, k is defined as the fundamental solution of A^*A. The RKHS norm becomes

$$\|Au\|^2 = \langle Au, Au \rangle = \langle u, A^*Au \rangle = \langle u, K^{-1}u \rangle =: \|u\|_k^2, \tag{2.47}$$

where $K = (A^*A)^{-1}$ is a covariance operator with $\langle \delta_x, K\delta_y \rangle = k(x,y)$. Smoothness of u can be achieved by defining u in the basis of the RKHS, $u \in \mathcal{H}_k$. Thus, each u satisfies the regularity of A. The new minimization functional to registration becomes

$$\underset{u \in \mathcal{H}_k}{\arg\min}\, \mathcal{D}[I_R, I_T, u] + \eta \|u\|_k^2, \tag{2.48}$$

where u is in the RKHS $\mathcal{H}_k$ and the regularity of u is fully defined by k. The following valuable fact makes this functional very important. It is guaranteed that a minimizer $\hat{u}$ of Equation 2.48 can be found in the parameters $\hat{\alpha}_i$, which follows from the representer theorem, see proofs in [85, 92]. Furthermore, it states that

$$\langle u, K^{-1}u \rangle = \langle \boldsymbol{\alpha}, K\boldsymbol{\alpha} \rangle. \tag{2.49}$$

Hence, a (local) solution is found applying gradient descent on the parameters.

The difference between the variational and the RKHS approach becomes apparent if for both approaches we plug-in the Gaussian kernel k_G. In the variational approach the Gaussian kernel implies that $A = -\nabla$ *resp.* $A^*A = -\Delta$ as it is the fundamental solution to the homogeneous

version of Equation 2.44. In the RKHS framework the kernel k defines the RKHS $\mathcal{H}_k$ over functions u which all satisfy the regularity of A where $A^*A = K^{-1}$. In the case of the Gaussian kernel, A is a linear combination of differential operators with even orders (*cf.* Steinke and Schölkopf [92]).

PROBABILISTIC INTERPRETATION The minimization problem in Equation 2.48 can be also interpreted as the MAP estimation problem of $\arg\max_u p(u)p(I_T|I_R, u)$, where $p(u) \propto \exp(-\mathcal{R}[u])$ is a Gaussian process prior and $p(I_T|I_R, u) \propto \exp(-\lambda^{-1}\mathcal{D}[I_R, I_T, u])$ is the likelihood (*cf.* Wahba [101]). A MAP solution can be found by solving the minimization problem (2.48) in the RKHS $\mathcal{H}_k$. The main advantage of using Gaussian processes to model the deformations lies in its flexibility. It will turn out later in this work (in Chapter 3), that the mean and the kernel function can be estimated from samples to obtain a statistical shape model constraint. Furthermore, generic kernel functions can be combined with a statistical shape model kernel.

In the following, we will spatially discretize Equation 2.48. This will result in a minimization scheme, which has a complexity of $\mathcal{O}(N^2)$ kernel evaluations, where N is the number of discretized points. Using a low-rank Gaussian process model, introduced in the following Section 2.4.2, the minimization of the functional becomes computationally feasible.

MATRIX-VALUED KERNELS Using an RKHS, not only scalar-valued (Section 2.2.4) but also vector-valued functions as *e.g.* displacement fields can be modeled. In this case, the kernel function becomes a matrix-valued function $k(x, y) : \Omega \times \Omega \to \mathbb{R}^{d\times d}$, with $k(x, y) = E[(u(x) - \mu)(u(y) - \mu)^*]$. The most simple case of matrix-valued kernel functions arises when the output dimensions are assumed to be uncorrelated. In this case, a matrix-valued kernel function k can be constructed from scalar-valued kernel functions κ by setting

$$k(x, y) = \mathbf{I}_d\kappa(x, y). \tag{2.50}$$

While vector-valued Gaussian processes seem like an extension of the theory, it can be shown that they can be reduced to the scalar case [43]. Thus, all known results from real Gaussian processes carry over to this more general setting. For the discretization of the RKHS minimization, which will follow next, we use matrix-valued kernels to express the vector-valued Gaussian process for the registration.

DISCRETIZATION Applied the RKHS optimization framework to registration, spatial discretization is needed. If the image domain Ω is sampled uniformly at $N \in \mathbb{N}$ points $\{x_i\}_{i=1}^N$, $u \in \mathcal{H}_k$ (including the mean μ) becomes

$$u(x) = \mu(x) + \sum_{i=1}^{N} k(x_i, x)\boldsymbol{\alpha}_{x_i} := \mu(x) + \mathbf{K}_x^T \boldsymbol{\alpha}, \tag{2.51}$$

where $\boldsymbol{\alpha}_{x_i} \in \mathbb{R}^d$, $\mathbf{K}_x = (k(x, x_i))_{i=1}^N \in \mathbb{R}^{Nd \times d}$ and $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_{x_i})_{i=1}^N \in \mathbb{R}^{Nd}$. This can be directly plugged into the full objective functional (2.41)

$$\arg\min_{\boldsymbol{\alpha} \in \mathbb{R}^{Nd}} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(I_T(x_i), I_R(x_i + \mu(x_i) + \mathbf{K}_{x_i}^T \boldsymbol{\alpha})) + \eta \boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha}, \tag{2.52}$$

where $\mathcal{L}$ is a loss function and $\mathbf{K} \in \mathbb{R}^{Nd \times Nd} | \mathbf{K}_{ij} = k(x_i, x_j) \in \mathbb{R}^{d \times d}$ is the kernel matrix. Equation 2.52 can be minimized using standard gradient descent based optimization (see Section A.3).

2.4.2 *Low-Rank GP Model*

The minimization of Equation 2.52 requires $\mathcal{O}(N^2)$ kernel evaluations. Especially in 3D, this becomes computationally demanding, since N becomes usually large. The goal is now, to express the terms in Equation 2.52 involving k, by an approximated and feasible representation of $\tilde{k}$. We perform principal component analysis on $\mathbf{K}$ to find an orthogonal basis expansion of k, such that

$$\arg\min_{\boldsymbol{\alpha} \in \mathbb{R}^{Nd}} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(I_T(x_i), I_R(x_i + \mu(x_i) + \mathbf{Q}\boldsymbol{\alpha})) + \eta \|\boldsymbol{\alpha}\|^2, \tag{2.53}$$

where $\mathbf{Q}$ is an orthogonal basis matrix.

Following the Mercer's theorem (see *e.g.* Rasmussen [77]), a kernel k has an expansion in terms of a orthonormal set of basis functions

$$k(x, y) = \sum_{i=1}^{\infty} \lambda_i \phi_i(x) \phi_i^*(y), \tag{2.54}$$

where (λ_i, ϕ_i) are the eigenvalue/eigenfunction pairs of the integral operator $\int_\Omega k(x, y) f(y) \mathrm{d}y$. A Hilbert space can be defined, by linear com-

binations of the eigenfunctions ϕ. Thus, each function $u = \sum_{i=1}^{\infty} \phi_i \boldsymbol{\alpha}_i$ with $\sum_{i=1}^{\infty} \frac{\boldsymbol{\alpha}_i^2}{\lambda_i} < \infty$. The RKHS norm becomes

$$\|u\|_k^2 = \sum_{i=1}^{\infty} \frac{\boldsymbol{\alpha}_i^2}{\lambda_i}, \tag{2.55}$$

and penalizes the eigenfunction components corresponding to small eigenvalues particularly strong. Now, each deformation in the space modeled by the Gaussian process $\mathcal{GP}(\mu, k)$ can be written as a sum $u(x) = \mu(x) + \sum_{i=1}^{\infty} \boldsymbol{\alpha}_i \phi_i(x)$. To get rid of the infinite sum, and to actually compute ϕ_i we propose to perform a low-rank approximation of the kernel k. We truncate the infinite sum to a finite sum over the main n basis functions ϕ_i, having the largest λ_i. If the eigenvalue spectrum of k decreases fast enough only few basis functions already lead to a good approximation.

THE NYSTRÖM APPROXIMATION Using the Nyström approximation [44, 77], the basis functions ϕ in the Mercer expansion can be computed. Let $N \in \mathbb{N}$ points $X_N = \{x_1, \ldots, x_N\}, x_l \in \Omega$ be sampled uniformly on the image domain Ω. The computation of ϕ_i becomes an eigenvalue problem:

$$\lambda_i \phi_i(x) = \int_{\Omega} k(x, y)\phi_i(y)\mathrm{d}y \approx \frac{1}{N} \sum_{l=1}^{N} k(x, x_l)\phi_i(x_l) \tag{2.56}$$

which results in a matrix eigenvalue problem

$$\tilde{\lambda}_i \mathbf{u}_i = \mathbf{K}\mathbf{u}_i, \tag{2.57}$$

where $\mathbf{K}_{j,l} = k(x_j, x_l)$, $\mathbf{u}_i$ denotes the i-th eigenvector and $\tilde{\lambda}_i$ the corresponding eigenvalue. $\tilde{\lambda}_i$ can directly be used as an approximation of the real eigenvalue λ_i, while the eigenfunction ϕ_i is approximated using

$$\tilde{\phi}_i(x) = \frac{\sqrt{N}}{\tilde{\lambda}_i} \mathbf{K}_{X_N}(x)^T \mathbf{u}_i \approx \phi_i(x), \tag{2.58}$$

with $\mathbf{K}_{X_N}(x) = (k(x_1, x), \ldots, k(x_N, x)) \in \mathbb{R}^{Nd \times d}$.

In medical image registration, N is usually large, *i.e.* beyond 1 million. Therefore, deriving the eigenvectors $\mathbf{u}$ is computationally demanding. However, the Nystöm method allows to solve the eigen system on a smaller sample size $X_M = \{x_1, \ldots, x_M\}, x_q \in \Omega$ and *extend* it to

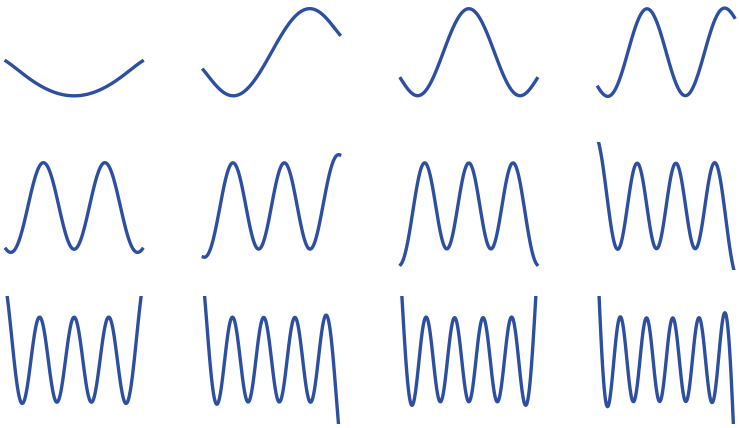


Figure 13: The first 12 leading eigenfunctions of a low-rank Gaussian $\sigma_G = 1, N = 200, x_i \in [-5, 5]$.

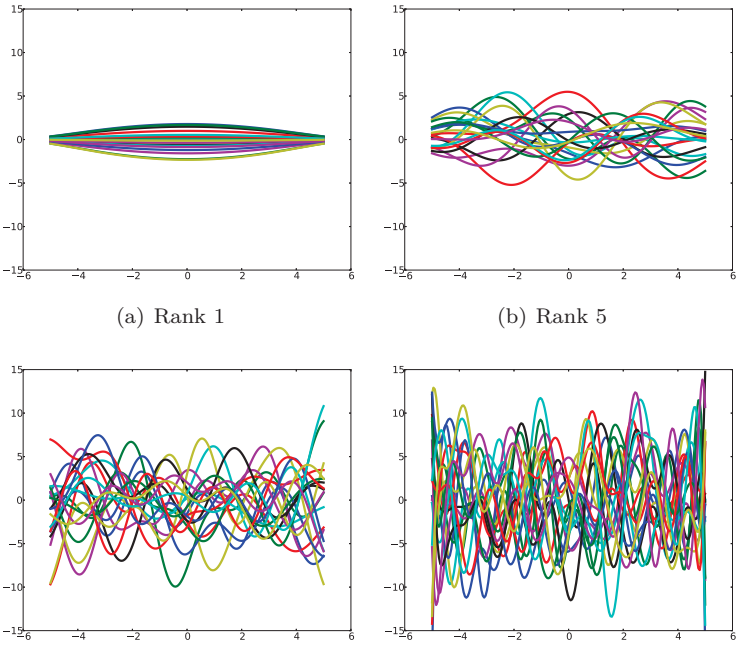
the remaining higher resolution sample set [44]. The estimates of the eigenvalue/eigenvector pairs become

$$\hat{\lambda}_i = \frac{N}{M}\tilde{\lambda}_i^M \qquad \hat{\phi}_i(x_l) = \sqrt{\frac{M}{N}}\frac{1}{\tilde{\lambda}_i^M}\mathbf{K}_{X_M}(x_l)^T\mathbf{u}_i^M, \tag{2.59}$$

where $\mathbf{K}_{X_M}(x_l) = (k(x_1, x_l), \ldots, k(x_M, x_l)) \in \mathbb{R}^{Md\times d}$. Thus, $\hat{\mathbf{u}}_i = (\hat{\phi}_i(x_l))_{l=1}^N$ can be precomputed to using it in the registration.

If the main l singular vectors of the subsampled kernel matrix $\mathbf{K}_M$ are approximated using the randomized SVD method of Halko et al. [40], the computational complexity for the approximation is $\mathcal{O}(2Ml^2)$.

LOW-RANK GP REGISTRATION With the help of the Nyström approximation, a low-rank approximation $\tilde{k}$ of the kernel k can be computed. This allows to construct a finite generative linear model of deformations u. Let u be spatially discretized on $N \in \mathbb{N}$ points $\mathcal{X} = \{x_i\}_{i=1}^N$



(a) Rank 1 (b) Rank 5

(c) Rank 10 (d) Rank 20

Figure 14: In this figure, 20 samples of the low-rank Gaussian process prior are shown, approximated with 1 (a), 5 (b), 10 (c) and 20 (d) eigenfunctions.

in the image domain. And let us sample the mean μ and the leading n basis functions $\hat{\phi}$ at these points as well

$$\boldsymbol{\mu} = (\mu(x_i))_{i=1}^{N} \in \mathbb{R}^{Nd} \tag{2.60}$$

$$\mathbf{Q} = (\mathbf{q}_1, \cdots, \mathbf{q}_n) \in \mathbb{R}^{Nd \times n}, \mathbf{q}_j = \left(\sqrt{\hat{\lambda}_j} \hat{\phi}_j(x_i) \right)_{i=1}^{N} \in \mathbb{R}^{Nd}. \tag{2.61}$$

The coefficient vector $\boldsymbol{\alpha}$ is normally distributed $\mathcal{N}(0, \mathbf{I}_n)$, and the deformations expressed by the model

$$u(\boldsymbol{\alpha}) = \boldsymbol{\mu} + \mathbf{Q}\boldsymbol{\alpha} \tag{2.62}$$

are distributed according to $\mathcal{N}(\boldsymbol{\mu}, \mathbf{Q}\mathbf{Q}^T)$.

This formulation allows to reformulate the objective functional for registration (2.48) to a parametric minimization functional

$$\arg\min_{\boldsymbol{\alpha} \in \mathbb{R}^n} \mathcal{D}[I_R, I_T, u_{\boldsymbol{\alpha}}] + \eta \|\boldsymbol{\alpha}\|^2, \tag{2.63}$$

with

$$u_{\boldsymbol{\alpha}}(x) = \boldsymbol{\mu}_x + \mathbf{Q}_x \boldsymbol{\alpha} \tag{2.64}$$

where the subscript indicates the i-th entry which corresponds to x, *resp.* the row-wise sub-matrix from row i to $i + d$.

Applying the Gaussian kernel k_G and a zero mean μ_0, for instance, registration is finally performed by minimizing Equation 2.63 using *e.g.* gradient descent based optimization methods. In Figure 13, approximated basis functions of the kernel k_G (in 1D) are shown. Since, we constructed a generative model $\mathcal{N}(\mathbf{0}, \mathbf{Q}\mathbf{Q}^T)$ out of the Gaussian process $\mathcal{GP}(\mu_0, k_G)$, samples can be drawn from the low-rank Gaussian process prior. In Figure 14, samples of the same 1D low-rank Gaussian process are plotted using different number n of approximated basis functions $\tilde{\phi}_i$.

2.4.3 *Landmark Posterior Model*

Using the Nyström approximation, the orthogonal basis functions of the Gaussian process prior k can be approximated. The obtained low-rank Gaussian process $\mathcal{GP}(\mu, \tilde{k})$ can be discretized on a finite domain resulting in a normal distribution of deformations $u \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{Q}\mathbf{Q}^T)$. Similar to the posterior Gaussian process derived in Section 2.2.4, a conditional distribution in finite domain given landmark displacements can be derived in closed form. The eigenvalue/eigenvector pairs of the new conditional model can be efficiently computed as well. Thus, the resulting linear model, conditioned on the landmarks,

$$u_c(\boldsymbol{\alpha}) = \boldsymbol{\mu}_c + \mathbf{Q}_c \boldsymbol{\alpha} \tag{2.65}$$

can conveniently integrated into the minimization functional for registration Equation 2.63. In the following, for the interested reader, we will provide the corresponding formulas to calculate the conditional model. We additionally refer to Albrecht et al. [9] for a deeper discussion about posterior models and to Bishop et al. [15], Chapter 2.3 for a detailed derivation of the closed form solutions.

CONDITIONAL MODEL Lets assume, p landmarks are given $X = \{(x_1, y_1), \ldots, (x_p, y_p)\} \subset \mathcal{X} \times \mathbb{R}^d$. Additionally, let assume that $y \sim \mathcal{N}(u(x), \sigma^2 \mathbf{I}_d)$ are noisy observations of u. Under the given distribution $\mathcal{N}(\boldsymbol{\mu}, \mathbf{Q}\mathbf{Q}^T)$

$$u_X(\boldsymbol{\alpha}) = \boldsymbol{\mu}_X + \mathbf{Q}_X \boldsymbol{\alpha} + \epsilon, \tag{2.66}$$

where $\boldsymbol{\mu}_X = (\boldsymbol{\mu}_{x_i})_{i=1}^p \in \mathbb{R}^{pd}$, $\mathbf{Q}_X = (\mathbf{Q}_{x_i})_{i=1}^p \in \mathbb{R}^{pd \times n}$, $x_i \in X$. Let us infer the distribution $p(\boldsymbol{\alpha}|u_X)$ and $p(u|u_X)$

$$p(\boldsymbol{\alpha}|u_X) = \mathcal{N}(\boldsymbol{\eta}, \boldsymbol{\Lambda}) \tag{2.67}$$

$$\boldsymbol{\eta} = (\mathbf{Q}_X^T \mathbf{Q}_X + \sigma^2 \mathbf{I}_n)^{-1} \mathbf{Q}_X^T (\mathbf{u}_X - \boldsymbol{\mu}_X) \tag{2.68}$$

$$\boldsymbol{\Lambda} = \sigma^2 (\mathbf{Q}_X^T \mathbf{Q}_X + \sigma^2 \mathbf{I}_n)^{-1}, \tag{2.69}$$

$$p(u|u_X) = \mathcal{N}(\boldsymbol{\mu} + \mathbf{Q}\boldsymbol{\eta}, \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^T) =: \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) \tag{2.70}$$

$$\boldsymbol{\mu}_c = \boldsymbol{\mu} + \mathbf{Q}(\mathbf{Q}_X^T \mathbf{Q}_X + \sigma^2 \mathbf{I}_n)^{-1} \mathbf{Q}_X^T (\mathbf{u}_X - \boldsymbol{\mu}_X) \tag{2.71}$$

$$\boldsymbol{\Sigma}_c = \sigma^2 \mathbf{Q} (\mathbf{Q}_X^T \mathbf{Q}_X + \sigma^2 \mathbf{I}_n)^{-1} \mathbf{Q}^T. \tag{2.72}$$

EIGEN ANALYSIS OF THE POSTERIOR MODEL If the number of spatially sampled points N is high, $\boldsymbol{\Sigma}_c = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^T \in \mathbb{R}^{Nd \times Nd}$ gets very large and cannot be stored in memory anymore. Therefore, to get the eigenvectors for the conditional model a detour is made. Let $\mathbf{D} \in \mathbb{R}^{n \times n}$ be a diagonal matrix with $\mathbf{D}_{ii} = \sqrt{\hat{\lambda}_i}$. Further, let $\mathbf{U} = \mathbf{Q}\mathbf{D}^{-1}$. Then,

$$\boldsymbol{\Sigma}_c = \mathbf{U} \underbrace{\sigma^2 \mathbf{D} (\mathbf{Q}_X^T \mathbf{Q}_X + \sigma^2 \mathbf{I}_n)^{-1} \mathbf{D}}_{\boldsymbol{\Sigma}_s} \mathbf{U}^T. \tag{2.73}$$

If eigen analysis is performed on $\boldsymbol{\Sigma}_s$, it decomposes into $\mathbf{U}_s \mathbf{D}_s^2 \mathbf{U}_s^T$, where $\mathbf{U}_s$ is the orthonormal basis matrix of $\boldsymbol{\Sigma}_s$ and $\mathbf{D}_s^2$ is a diagonal matrix with the corresponding variances on the diagonal. It follows that $\boldsymbol{\Sigma}_c = (\mathbf{U}\mathbf{U}_s)\mathbf{D}_s^2(\mathbf{U}_s\mathbf{U})^T$ and therefore

$$\mathbf{Q}_c = (\mathbf{U}\mathbf{U}_s)\mathbf{D}_s. \tag{2.74}$$

The conditional model can be represented as $u_c(\boldsymbol{\alpha}) = \boldsymbol{\mu}_c + \mathbf{Q}_c \boldsymbol{\alpha}$ and directly be used in the minimization functional for registration Equation 2.63.

DISCUSSION

So far, we have introduced the regularized minimization functional Equation 2.5 for image registration. We have first focused on the variational

Variational Registration

approach to minimize the functional. With the linearization of the corresponding system of semilinear partial differential equations and having given its fundamental solution, an iteration scheme leads to a local solution. We extended this framework by kernel functions which are nonseparable *resp.* nonstationary. This enables to apply a kernel, which incorporates prior landmark displacements.

RKHS Registration

In contrast to the variational approach, we introduced the RKHS framework for registration. The registration functional Equation 2.5 is directly minimized in a reproducing kernel Hilbert space, where the smoothness is defined by the corresponding kernel function k. To make the minimization feasible even for large images (3D CT images), we have proposed a low-rank approximation of the kernel using the Nyström approximation. We derived a posterior low-rank Gaussian process prior, conditioned on landmark displacements, to enable hybrid landmark and image registration.

Difference between two Approaches

Applied the same operator A to each of the presented approaches, not the very same problem is minimized. In the variational approach, no assumptions about the displacement field u are made when calculating the forces. The fundamental solution k of the time-dependent PDE $\frac{\partial}{\partial t} - A^*A$ induces smoothness to the solution u. Contrary, in the RKHS approach, the fundamental solution k of A^*A defines the smoothness of the solution which is sought, to minimize the functional (Equation 2.5). In the experiments later in this work (Chapter 4), the performance of the two approaches is evaluated and compared. However, we now roughly conclude the different properties of the two approaches, to give an intuition about how they can be applied.

Spatial Support

The efficiency of the variational approach strongly depends on the properties of the kernel function. If k is stationary, the extra memory needed can be neglected, and the optimization can be performed very fast (*cf.* Section 4.1). In the case of a nonstationary kernel, the memory consumption and the computational costs depend on the spatial variation of the kernel. For the landmark kernel, this keeps within reasonable bounds. However, for kernels having a large support, the computational costs explode.

Decreasing Eigenvalue Spectrum

The RKHS registration impresses with its flexibility to construct a prior. Furthermore, samples from the prior can be drawn, which can be used to visualize the designed prior. The extra memory and the computational costs for the minimization depend on the eigenvalue spectrum of the kernel. If it decreases fast, the extra memory needed to store the eigenvectors is low, while the computational costs for minimization are low as well. This is the case, if only small variations in the deformations

are allowed. On the other hand, if a large variation is allowed, *e.g.* a small σ_G of the Gaussian kernel, a large number of eigenfunctions have to be approximated using a lot of sampled points in the Nyström approximation. This requires more computational power to decompose the subsampled kernel matrix and a lot of memory to store basis matrix.

CONCLUSION In this chapter, we thoroughly studied two approaches for registration. By providing a kernel function, in both approaches a prior about admissible deformations is defined. In particular, we incorporate landmarks as prior deformations. This allows to locally improve the registration.

In the next chapter, the RKHS registration framework is used to apply a statistical shape model to robustly segment medical images. A kernel function is constructed which captures the covariance of already registered samples. Thus, statistical model fitting can be viewed as registration with an empirical kernel function.

3 STATISTICAL MODEL

In the previous chapter, two strategies to optimize the registration functional have been presented. The goal is to express a target image through a reference image by finding a smooth mapping from the reference to the target. A generic smoothness prior, which penalizes high derivatives for instance, ensures the resulting transformation to be smooth. But, this approach is often too generic, and does not incorporates prior knowledge about the object class when the registration is performed. However, in many applications, sample data is available, which can be used for constructing a shape prior over an object class. Moreover, beside an empirical shape model, the appearance can be learned from sample data as well, to using it in the data term *i.e.* in the matching criterion of the fitting procedure.

Learning from samples

In this chapter, we introduce the idea of statistical shape models and how they fit in the low-rank Gaussian process registration framework presented above. Additionally, we examine in more detail the matching criterion *i.e.* the image to image metric, when fitting a statistical shape model to unobserved data. Having available the sample images, which are also used to build the statistical shape model, we present a nonparametric statistical appearance model, which makes it possible to robustly fit the statistical shape model to a target image.

3.1 STATISTICAL SHAPE MODEL

Recalling the registration problem, discussed in the previous section, a target image was expressed by a reference image, by finding a smooth transformation from one to another. The space of admissible deformations, was restricted to deformations which are smooth in a rather generic sense, of small gradients for instance. Regarding the scenario, where some sample images and corresponding object label maps of the same object class are given, the space of admissible deformations could be further restricted. The sample images can be registered in a simplified setting since ground truth labels are given. The resulting transformations are used as a prior in the registration of unseen images. The assumption, which is usually made, is that the space of transformations which are considered is a linear space in $\mathbb{R}^p$ (p = number of samples)

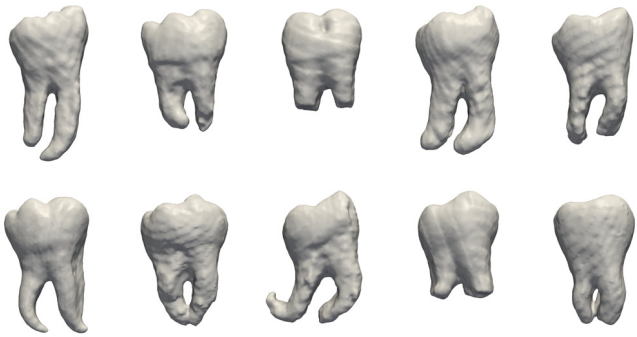


Figure 15: 10 sample tooth shapes.

and that an instance $s \in \text{span}(\mathcal{S})$ can be formed by any linear combination of a set of sample transformations $\mathcal{S}$. Thus, registration becomes a search for the coefficients of the linear combination which leads to the best match. This is also referred to as model fitting.

Let $p \in \mathbb{N}$ sample displacement fields $\mathcal{S} = \{s_i\}_{i=1}^{p}$ be given. We restate the minimization functional for registration

$$\arg\min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \mathcal{D}[I_R, I_T, \sum_{i=1}^{p} \alpha_i s_i] + \mathcal{R}[\boldsymbol{\alpha}]. \tag{3.1}$$

The choice of the regularization $\mathcal{R}[\boldsymbol{\alpha}]$ defines the prior over the samples. In Figure 15, samples of tooth shapes are shown. Note that the samples s_i are displacement fields which are applied to a reference shape. These displacement fields are the outcome of the registration of the reference to the sample tooth shapes.

GAUSSIAN PROCESS PRIOR Lets assume the sample transformations are a Gaussian process $s \sim \mathcal{GP}(\mu, k)$. Since μ *resp.* k are not known, they are estimated from the given samples in $\mathcal{S}$. From the definition of Gaussian processes follows, that each finite set of samples from a given Gaussian process are jointly normal distributed. As such, the empirical

mean $\mu_{\mathcal{S}}(x)$ and the empirical covariance $k_{\mathcal{S}}(x, y)$ of this normal distribution $\mathcal{N}(\mu_{\mathcal{S}}, k_{\mathcal{S}})$ are given as follows

$$\mu_{\mathcal{S}}(x) = \frac{1}{p} \sum_{i=1}^{p} s_i(x) \tag{3.2}$$

$$k_{\mathcal{S}}(x, y) = \frac{1}{p-1} \sum_{i=1}^{p} (s_i - \mu_{\mathcal{S}})(x) \otimes (s_i - \mu_{\mathcal{S}})(y), \tag{3.3}$$

where $\otimes$ is the outer product.

Following the low-rank Gaussian process model of Chapter 2, we are interested in a parametrization, where the parameters $\boldsymbol{\alpha}$ are normally distributed according to $\mathcal{N}(0, \mathbf{I}_p)$, such that a model instance becomes

$$s_{\boldsymbol{\alpha}}(x) = \mu_{\mathcal{S}}(x) + \sum_{i=1}^{p} \alpha_i \psi_i(x), \tag{3.4}$$

where ψ_i are orthogonal basis functions. The computation of ψ_i becomes an eigenvalue problem:

$$\lambda_i \phi_i = \int_{\Omega} k_{\mathcal{S}}(x, y) \phi_i(y) dy, \tag{3.5}$$

where $\psi_i = \sqrt{\lambda_i} \phi_i$. Spatially discretized with $\{x_l\}_{l=1}^{N}, x_l \in \Omega, N \in \mathbb{N}$, Equation 3.5 results in a matrix eigenvalue problem

$$\tilde{\lambda}_i \mathbf{u}_i = \mathbf{K}_{\mathcal{S}} \mathbf{u}_i, \tag{3.6}$$

where $\mathbf{K}_{\mathcal{S}ij} = k_{\mathcal{S}}(x_i, x_j)$, $\mathbf{u}_i$ denotes the i-th eigenvector and $\tilde{\lambda}_i$ the corresponding eigenvalue. Like in Chapter 2, the approximated eigenfunctions $\tilde{\phi}_i$ could be derived using the Nyström approximation. However, in this case the rank of the kernel $k_{\mathcal{S}}$ is exactly $p - 1$ compared to the Gaussian kernel k_G which has infinite rank. Thus, in this case, the rank is much smaller than the number of sampled points N. Therefore, instead of performing the Nyström approximation, principal component analysis (PCA) is more efficient.

PRINCIPAL COMPONENT ANALYSIS Recall the Mercer theorem for $k_{\mathcal{S}}$, which states that each kernel can be expanded as follows,

$$k_{\mathcal{S}}(x, y) = \sum_{i=1}^{p} \lambda_i \phi_i(x) \phi_i^*(y), \tag{3.7}$$

where ϕ_i are orthonormal functions (since only p samples are given, the sum goes from 1 to p). In the finite domain, this means, $\mathbf{K}_\mathcal{S}$ can be decomposed as follows

$$\mathbf{K}_\mathcal{S} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T, \tag{3.8}$$

where $\mathbf{U} \in \mathbb{R}^{Nd \times p}$ is an orthonormal matrix $\mathbf{U}^T\mathbf{U} = \mathbf{I}_p$ and $\mathbf{\Lambda}$ is a diagonal matrix with $\mathbf{\Lambda}_{ii} = \lambda_i$. Multiplied by $\mathbf{U}$, Equation 3.8 becomes

$$\mathbf{K}_\mathcal{S}\mathbf{U} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T\mathbf{U} = \mathbf{U}\mathbf{\Lambda}, \tag{3.9}$$

which corresponds to the matrix eigenvalue problem Equation 3.6 for the i-th column $\mathbf{u}_i$ of $\mathbf{U}$.

Let $\mathbf{S} \in \mathbb{R}^{Nd \times p}$ be a matrix where the columns are the vectorized mean free sample displacement fields $s_i - \mu_\mathcal{S}$, sampled a N points. The empirical covariance matrix, which is equivalent to the kernel matrix $\mathbf{K}_\mathcal{S}$ is

$$\mathbf{K}_\mathcal{S} = \frac{1}{p-1}\mathbf{S}\mathbf{S}^T \in \mathbb{R}^{Nd \times Nd}. \tag{3.10}$$

Because $\mathbf{K}_\mathcal{S}$ and the smaller matrix $\mathbf{C}_s = \frac{1}{p-1}\mathbf{S}^T\mathbf{S}$ share the first $p-1$ main eigenvalues (see more details in Section A.4), the eigenvectors of $\mathbf{K}_\mathcal{S}$ with non-zero eigenvalues can be derived by

$$\mathbf{U}_\mathcal{S} = \mathbf{S}\mathbf{V}_s\mathbf{D}_s^{-1} \quad \text{with} \quad \mathbf{C}_s = \mathbf{U}_s\mathbf{D}_s^2\mathbf{V}_s^T. \tag{3.11}$$

The linear generative model, which spans the samples in $\mathcal{S}$ with $\boldsymbol{\alpha} \sim \mathcal{N}(0, \mathbf{I}_p)$ finally becomes

$$s(\boldsymbol{\alpha}) = \boldsymbol{\mu}_\mathcal{S} + \mathbf{Q}_\mathcal{S}\boldsymbol{\alpha}, \tag{3.12}$$

where $\boldsymbol{\mu}_\mathcal{S} \in \mathbb{R}^{Nd}, \boldsymbol{\mu}_{\mathcal{S}i} = \mu_\mathcal{S}(x_i)$ and $\mathbf{Q}_\mathcal{S} = \mathbf{U}_\mathcal{S}\mathbf{D}_s$. The minimization functional for model fitting, using a Gaussian prior on the given samples $\mathcal{S}$ becomes

$$\underset{\boldsymbol{\alpha} \in \mathbb{R}^p}{\arg\min}\, \mathcal{D}[I_R, I_T, \boldsymbol{\mu}_\mathcal{S} + \mathbf{Q}_\mathcal{S}\boldsymbol{\alpha}] + \eta\|\boldsymbol{\alpha}\|^2 \tag{3.13}$$

and has the form of a low-rank Gaussian process model of the previous Chapter 2, which can be minimized with the same methods. In Figure 16, the mean and the first 5 main variations of a tooth model are visualized.

Figure 16: In this figure the 5 main principal components are row-wise visualized. Upper row: $+1.5\sigma$, lower row: -1.5σ. The model mean is depicted in blue.

VARIATIONAL STATISTICAL SHAPE MODEL FITTING Applying the statistical model kernel $k_{\mathcal{S}}$ to the variational registration framework introduced in Chapter 2

$$u^{i+1} = \int_{\Omega} k_{\mathcal{S}}(\cdot, x)(u^i + f(u^i))\mathrm{d}x, \tag{3.14}$$

is computationally very expensive. That is, because $k_{\mathcal{S}}$ has "global" support, *i.e.* $k_{\mathcal{S}}(x, y)$ varies for each $y \in \Omega$ holding x fix. Furthermore, $k_{\mathcal{S}}$ spatially varies for each $x \in \Omega$, which exceeds the capabilities of the caching scheme.

3.1.1 *Extended Statistical Model Kernels*

Enhancing the Model Span

Since the true distribution among the considered object class is only approximated by a linear combination of a finite training set, parts, which cannot be explained by the model during model fitting, are biased by the model. While this property is actually desired when using statistical shape models, it might be too restrictive. This is the case especially, if only few samples are given. A combination of a generic deformation model on top of the statistical shape model can be used to reduce such a bias. In the following, we present two different approaches where the model bias is reduced.

BIAS REDUCTION In the low-rank Gaussian process framework, the full prior is encoded in a single kernel function. We therefore propose to construct a kernel which allows a generic deformation model on top of the empirical shape model. This is achieved by combining the statistical model kernel with a Gaussian kernel

Combining Kernels

$$k_B = k_{\mathcal{S}} + \tau k_G, \tag{3.15}$$

where τ indicates the scale of the generic deformation. The bias reduced kernel k_B can be interpreted as follows. On top of a deformation which is regular in $k_{\mathcal{S}}$, we add a generic deformation, which is itself regular under k_G, but irregular under $k_{\mathcal{S}}$. The combined deformation thus is regular with respect to the bias reduced kernel k_B. Using the Nyström approximation, the combined kernel can be approximated to get the low-rank approximation of the corresponding Gaussian process model $\mathcal{GP}(\mu_{\mathcal{S}}, k_{\mathcal{S}} + \lambda k_G)$. The impact of the bias reduced model is evaluated in the experiments Chapter 4.

General kernel combinations, which result again in valid kernel functions are provided in Section A.2.

GEODESIC CORRELATION DAMPING Modeling the correlation between parts of the shape in terms of the Gaussian kernel k_G means that points which are spatially close are strongly correlated, while points having a large Euclidean distance are barely correlated. Considering the tooth shape, and in particular the tooth's root tips, this model does not intuitively hold. The root tips might be spatially quite close, while they do not have a lot to do with each other. However, incorporating the geodesic distance on the surface of the object to model the point-to-point correlations is much more intuitive. We present a generic kernel, which is aware of the geodesic distance on the surface. For using such a kernel, in the low-rank Gaussian process framework presented in this work, we have to restrict the image domain to a surface domain.

Considering Topology

Let $L_R : \Omega \subset \mathbb{R}^3 \to \{0, 1\}$ be a reference label map on the image domain Ω. The boundary of the object is defined as the surface $\Gamma_R := \{x | x \in |\nabla L_R(x)| > 0\} \subset \mathbb{R}^3$. Further, let be $u_{\Gamma_R} : \Gamma_R \to \mathbb{R}^3$ a displacement field, defined on the domain Γ_R. Then, the deformation prior in the registration *resp.* statistical shape model fitting can be modeled by replacing the domain in the low-rank Gaussian process framework with Γ_R. The data term, *i.e.* the image distance, becomes a distance between the warped reference surfaces $\Gamma_R + u_{\Gamma_R}$ and a target surface Γ_T.

Object Boundary as Surface

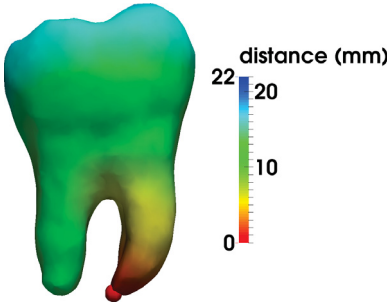


Figure 17: The geodesic shape distance starting from the right root tip is color coded on the tooth shape surface.

Let define a function $\psi_\Gamma : \Gamma_R \times \Gamma_R \to \mathbb{R}$, which returns the geodesic distance of two points on Γ_R[1]. Having further constructed a statistical surface model kernel $k_\mathcal{S}$, based on sample displacement fields $\mathcal{S}_\Gamma = \{s^i_{\Gamma_R}\}_{i=1}^p$, $k_\mathcal{S}$ can be combined with the following *geodesic kernel*[2]

$$k_\Gamma(x,y) = \exp\left(-\frac{\psi_\Gamma(x,y)^2}{\sigma^2}\right), \quad x,y \in \Gamma_R. \tag{3.16}$$

Euclidean Distance

The model bias reduction can be obtained by damping the correlation of a point-pair if they are geodesically wide apart

$$k_{B_\Gamma} = k_\mathcal{S} \cdot k_\Gamma. \tag{3.17}$$

Euclidean Distance

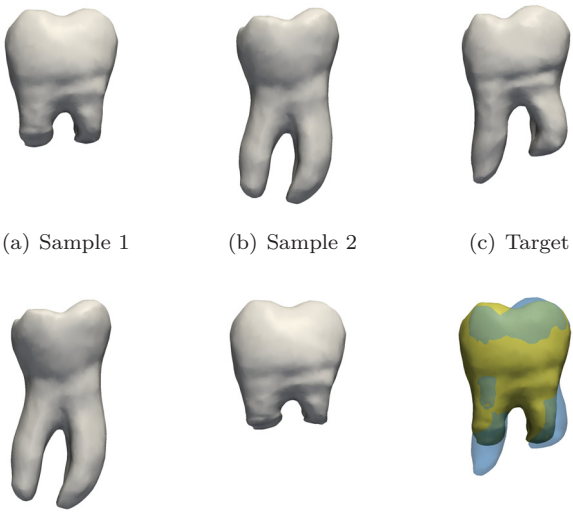
Technically, ψ_Γ could be implemented by pre-computing the shortest path distance of all point pairs on a discretized mesh of the surface Γ_R. Consequently, the Nyström approximation of k_{B_Γ} is also performed only on these discretized points. In Figure 17, the geodesic distance from the right root tip to all other points on the tooth surface is visualized in color.

To show the effect the geodesically damped shape model has, in Figure 18, a simple empirical shape model has been built using two sample tooth surfaces, one with underdeveloped roots and one with relatively long roots. The first and only main variation of this model is visualized as well. The model fit to the target tooth shape, which has different root

Geodesically Damped Surface Model

1 It could be implemented with the Floyd-Warshall algorithm Cormen et al. [27], Section 26.2.

2 ψ_Γ could also be squared.



(a) Sample 1 (b) Sample 2 (c) Target

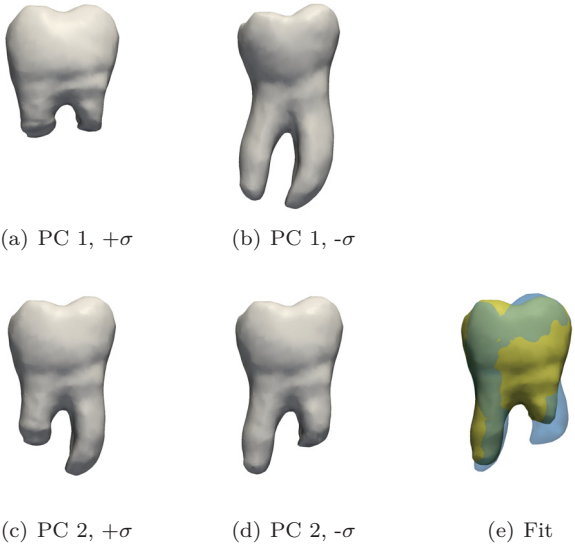
(d) PC 1, $+\sigma$ (e) PC 1, $-\sigma$ (f) Fit

Figure 18: (a) and (b) are the sample shapes to build the empirical model. (c) is the target to with the model is fitted. (d) and (e) visualize the first and only eigenmode $\pm$ 1 standard deviation. (f) shows the model fit in yellow and the again the target shape in blue.

length, is the best compromise with roots which are equally long. Additionally, the first two eigenmodes of a geodesically damped model have been approximated. The main variations are visualized in Figure 19. By adding this flexibility, the model is able to fit the target with different root lengths. In the experiments Chapter 4, we quantitatively evaluate the geodesically damped model using a surface to image similarity metric presented in the next section.

3.2 STATISTICAL APPEARANCE MODEL

In the previous section, the focus was on building a prior over the admissible deformations given a sample dataset. Sample displacement fields $\mathcal{S} = \{s_i\}_{i=1}^p$ have been given, which originate from the registration of the sample images $\mathcal{I} = \{I_i\}_{i=1}^p$ *resp.* the corresponding sample label maps $\mathcal{L} = \{L_i\}_{i=1}^p$ to a reference image I_R *resp.* reference label map



(a) PC 1, $+\sigma$ (b) PC 1, $-\sigma$

(c) PC 2, $+\sigma$ (d) PC 2, $-\sigma$ (e) Fit

Figure 19: In the first row the first eigenmode of the geodesically damped model is visualized $\pm$ 1 standard deviation. (c) and (d) show one further approximated eigenmode $\pm$ 1 standard deviation as well. (e) shows the model fit in yellow and again the target in blue.

L_R. In this section, we focus on the data term $\mathcal{D}$ of the minimization functional for model fitting Equation 3.13.

Point-to-point correspondence

In the registration as well as in the model fitting, the reference image I_R is used as an atlas. Having established point-to-point correspondence to a target image I_T, features such as object labels can be transfered form the reference to the target. This approach involves several problems. For instance, it is assumed, that for each point in the reference image there exists a corresponding point in the target. Lets look at the CBCT-tooth images considered in this work (see three examples in Figure 20). At least in the background, defining correspondence is difficult. Examining the tooth shape in more detail, we conclude, that even inside the tooth object, the point-wise correspondence assumption is incongruous. In Figure 21, examples of the nerve structure of the inside of the tooth, called pulp, are visualized. Despite the tooth has only two roots, the pulps can have a different number of tracts. Additionally, the several hunches make the registration of the pulp difficult.

Figure 20: Sample image slices of CBCT images where the wisdom tooth shape is outlined in *blue*.

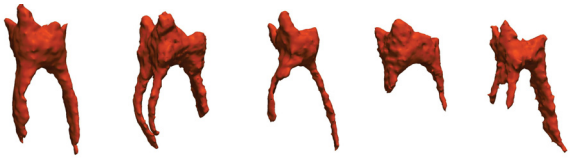
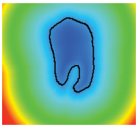


Figure 21: Sample pulp shapes.

These correspondence problems occur not only in the model fitting, but already in the registration of the sample images. To simplify the registration, instead of the CBCT images, distance maps of the given tooth label maps are registered (see Section 2.1.3). Thus, the background as well as the correspondence-less parts inside the object are omitted, and only the tooth surface is considered. However, the problems remain in the model fitting. It is difficult to find a reference, where the pulp structure inside the tooth and the background can be explained with such a statistical model.

Label map

To tackle the background problem there are different strategies. An easy way to relieve the misleading influence of the background is to simply mask it out. That means, only the image difference inside the object is considered. However, this again involves problems. Since only the object inside contribute to the matching error, the shrinkage of the object would give lower costs. Therefore, the background has to be included in the image metric.



Signed distance map

We propose to learn an appearance model, based on the given sample images $\mathcal{I}$, which can estimate the object *resp.* background probability of

a point in the target image I_T. The outcoming probability map will be used in the image metric presented in Section 3.2.2. The goal with this approach is to predict the background as accurate as possible, and to predict the pulp structure as object together with the dentin/enamel. Since we do not expect a perfect object/background prediction, a statistical shape model is fitted onto the probability map.

SEGMENTATION PROBLEM REVISED The segmentation problem can be formulated as follows. A reference label map $L_R : \Omega \subset \mathbb{R}^d \rightarrow \{0, 1\}$ is deformed by a parametric deformation model $\mathcal{M}[\boldsymbol{\alpha}](x) : \mathbb{R}^p, \mathbb{R}^d \rightarrow \mathbb{R}^d$, such that it matches the unknown target label map L_T, which assigns each point of a corresponding target image $I_T : \Omega \rightarrow \mathbb{R}$ as belonging to object or background. Since L_T is not known, the object probability of the target image is predicted. The prediction is represented by a probabilistic shape function $P_T : \Omega \rightarrow [0, 1]$, where $P_T(x)$ denotes the probability of a point $x \in \Omega$ being part of the object of interest. The minimization functional becomes

$$\arg\min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \mathcal{D}_P[P_T, L_R, \mathcal{M}[\boldsymbol{\alpha}]] + \eta \mathcal{R}[\boldsymbol{\alpha}]. \tag{3.18}$$

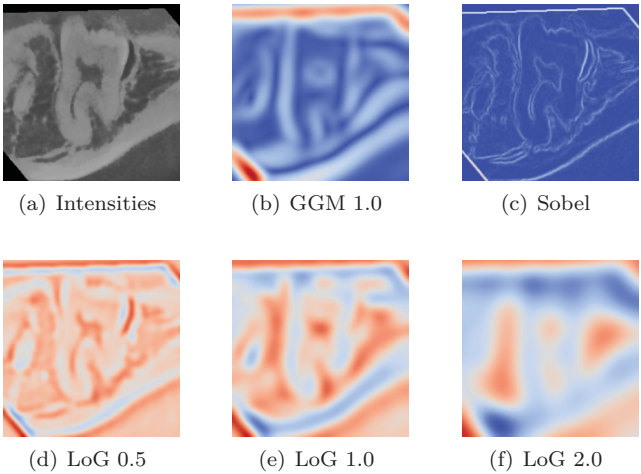
Instead of comparing intensity values, $\mathcal{D}_P$ quantifies the matching between the deformed reference label map and the predicted target probability map. Hence, all that is needed is an accurate object and background predictor for I_T.

3.2.1 *Random Forest Regression*

Because of the correspondence-less pulp structures, an appearance model is needed which is correspondence-less as well. We propose to learn the density $p(P_T|\mathcal{I}, \mathcal{L}, I_T)$ in a nonparametric and discriminative manner using random forest regression. Additional to the sample images $\{I_i\}_{i=1}^p$ and the corresponding ground truth label maps $\{L_i\}_{i=1}^p$, let further $F_i : \Omega \rightarrow \mathbb{R}^m$ be a feature image of I_i obtained by a general feature extraction method. Hence, for each point in the image, there is p-times an m-dimensional feature vector $F_i(x)$ with a corresponding binary label $L_i(x)$ indicating the point being object or background. With this training data, the object and background prediction can be learned by any regression method. Since there is a lot of training data, we propose to use random forest regression.

Random forest regression

If only the image intensities are considered as feature, the intensity distribution (histogram) is estimated. However, more advanced features



(a) Intensities (b) GGM 1.0 (c) Sobel

(d) LoG 0.5 (e) LoG 1.0 (f) LoG 2.0

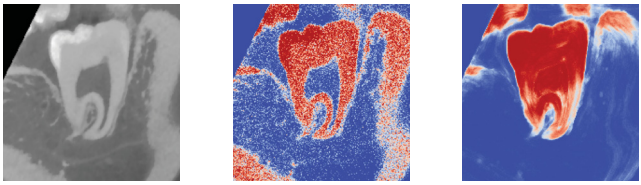
Figure 22: The following features are visualized: (a) intensity, (b) Gaussian gradient magnitudes with $\sigma = 1$, (c) Sobel edges, (d) Laplacian of Gaussians $\sigma = 0.5$, (e) Laplacian of Gaussians $\sigma = 1.0$ and (f) Laplacian of Gaussians $\sigma = 2.0$.

as *e.g.* Laplacian of Gaussians, Sobel edges or Harmonic Filters, incorporate neighborhood relations and are able to capture long range structural dependencies of different parts in the object. For instance, the pulp structure inside the tooth can be predicted as object as well. In Figure 22, an example of the different features are visualized. In Figure 23, a probability map based on the intensity values, and one based on the mentioned features are depicted.

3.2.2 *Probabilistic Similarity Metric*

The obtained probability map could be directly compared to the deformed label map by a squared loss for instance. However, we propose a more advanced distance metric, which also takes the boundary of the object into account.

In the following, we define a rather general (dis)similarity metric $\mathcal{D}_P$. Recall, that a global consistency in the object estimate is established by fitting the statistical shape model. At this stage however, the ob-



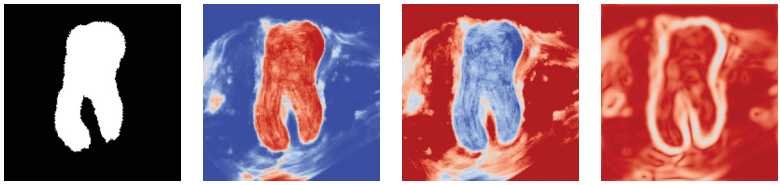
(a) CBCT image (b) PM intensity (c) PM feature

Figure 23: (a) Intensities of the CBCT image, (b) probability map obtained by random forest regression learned on solely the intensities, (c) probability map where various features have been used for training. *Red* corresponds to the object probability of 1, *blue* to 0.

ject and background probability for each point in the image domain is estimated separately and treated as independent. Hence, for the full object *resp.* background probability, we integrate over the logarithms of all object *resp.* background points in the image. Additionally, as object boundary term, a weighted total variation (TV) norm [79] is added to the metric. An edge indicator serves as the weighting. Note, that classically, the image gradients are used for the weighting in the TV norm. But, in our low quality images, especially in the region of the tooth's roots, it is difficult to estimate edges. Therefore, we use the object probability map for the edge indicator.

Let the deformed reference label map be denoted as $L_R(x+u(x))$ and the probability map for the object as P_T. The similarity measure is

$$\begin{aligned}\mathcal{D}_P[u] =& \lambda \int_\Omega \underbrace{-\log(P_T(x))L_R(x+u(x))\mathrm{d}x}_{\text{foreground}} \\ &+ \lambda \int_\Omega \underbrace{-\log(1-P_T(x))(1-L_R(x+u(x)))\mathrm{d}x}_{\text{background}} \\ &+ (1-\lambda) \int_\Omega \underbrace{\frac{1}{1+|\nabla P_T(x)|}|\nabla L_R(x+u(x))|\mathrm{d}x}_{\text{boundary}},\end{aligned} \tag{3.19}$$



(a) $L_R(x + u^*(x))$ (b) $P_T(x)$ (c) $P(x)$ (d) $Q(x)$

Figure 24: (a) is the reference label map, which is transformed with the ground truth displacement field, (b) depicts the probability map of the target image obtained by random forest regression, (c) is the simplified regional part $P(x)$ of the metric and (d) is the boundary part $Q(x)$. The interval $[0, 1]$ has been color coded with [*blue*,*red*] (except for the label map). The gradients have been calculated by Gaussian gradients with $\sigma = 0.3$.

where λ is a trade-off parameter between the regional and the boundary terms. The long expression can be simplified to

$$\begin{aligned}\mathcal{D}_P[u] =& \lambda \int_\Omega \log\left(\frac{1 - P_T(x)}{P_T(x)}\right) L_R(x + u(x))\mathrm{d}x \\ &+ (1 - \lambda) \int_\Omega \frac{1}{1 + |\nabla P_T(x)|} |\nabla L_R(x + u(x))|\mathrm{d}x.\end{aligned} \tag{3.20}$$

Finally, the first factor of the region part is referenced as $P(x)$ and the one of the boundary part to $Q(x)$

$$\begin{aligned}\mathcal{D}_P[u] =& \lambda \int_\Omega P(x) L_R(x + u(x))\mathrm{d}x \\ &+ (1 - \lambda) \int_\Omega Q(x) |\nabla L_R(x + u(x))|\mathrm{d}x.\end{aligned} \tag{3.21}$$

For illustration purposes, the individual terms and components appearing in this metric are visualized by an example in Figure 24 and Figure 25. The terms are shown for an optimal displacement field u^* and for a suboptimal one u'. One can see, that for the optimal displacement field, the regional part as well as the boundary part give low values (blue), while for the suboptimal one, they become more "red" *i.e.* higher values.

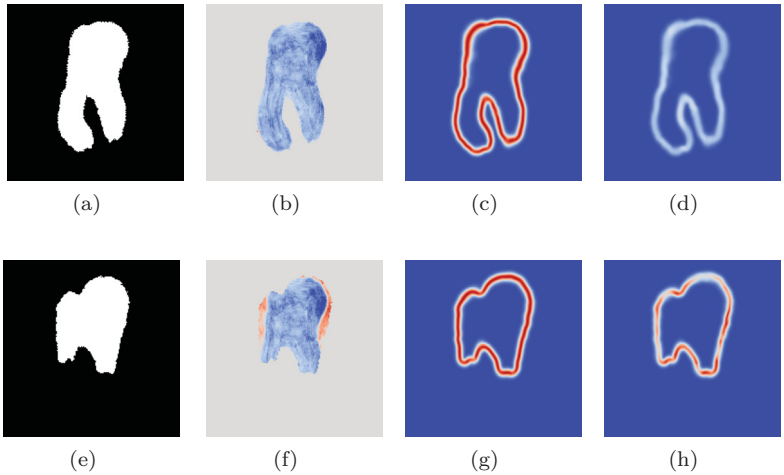


Figure 25: (a) is the reference label map which is deformed by the ground truth displacement field $L_R(x + u^*(x))$. In (b), the metric value of the regional part with this ground truth label map $P(x)L_R(x+u^*(x))$ is depicted. (c) is the gradient magnitude of the ground truth deformed reference label map $|\nabla L_R(x + u^*(x))|$. In (d), the metric value of the boundary part with the ground truth boundary $Q(x)|\nabla L_R(x + u^*(x))|$ is depicted.
(e) is the reference label map which is deformed by another displacement field $L_R(x + u'(x))$. In (f), the metric value of the regional part with this other label map $P(x)L_R(x + u'(x))$ is depicted. (g) is the gradient magnitude of the reference label map deformed by another displacement field $|\nabla L_R(x+u'(x))|$. In (h), the metric value of the boundary part with this other boundary $Q(x)|\nabla L_R(x+u'(x))|$ is depicted. The colors in the second column are coded as $[-4, 4]$ = [*blue*,*red*] and 0 = gray. In last two columns the colors are $[0, 1.4]$ = [*blue*,*red*].

The derivative of this metric[3] with respect to the displacement field is given as

$$\begin{aligned}\frac{d\mathcal{D}_P[u]}{du} =&\lambda \int_\Omega P(x)(\nabla L_R)(x+u(x))\mathbf{J}_u(x)\mathrm{d}x \\ &+ (1-\lambda)\int_\Omega Q(x)\frac{(\nabla L_R)(x+u(x))\mathbf{J}_u(x)}{|(\nabla L_R)(x+u(x))\mathbf{J}_u(x)|} \\ &\cdot \Big(\big((\nabla^2 L_R)(x+u(x))\mathbf{J}_u(x)\big)^T\mathbf{J}_u(x) \\ &\qquad + (\nabla L_R)(x+u(x))\mathbf{J}_u(x)\mathbf{H}_u(x)\Big)\mathrm{d}x,\end{aligned} \tag{3.22}$$

where $\mathbf{J}$ denotes the Jacobian. Since in the low-rank Gaussian process framework u is linear in the parameters, the Hessian $\mathbf{H}$ vanishes and the derivative can be efficiently computed.

3.2.3 *Shape to image metric*

The metric presented in the previous section is applied if a statistical deformation model is given. However, the regional part of the metric does not make sense if a statistical surface model is fitted to a target as it is the case for instance with the geodesic model. In this case, we propose the following surface to image metric which is only based on the boundary of the object

$$\mathcal{D}_\Gamma[\boldsymbol{\alpha}] = \int_{\Gamma_R} \frac{1}{1+|\nabla P_T(x+\mathcal{M}[\boldsymbol{\alpha}](x))|}\mathrm{d}x. \tag{3.23}$$

In Chapter 4, we will show experiments with this metric and the geodesic surface regularized model.

DISCUSSION

Combined Models

We introduced the standard approach to statistical shape model construction. To obtain the empirical basis, we provided the viewpoint of the previous introduced low-rank approximation, where an orthogonal basis expansion of a kernel is derived. Further, we presented two different approaches to relax the model bias. Both are based on kernel composition. In the first approach, a generic Gaussian kernel is added to the statistical model kernel. The second approach focuses on plain

3 Note: the following rule for differentiating is needed: $f(x) = |a|, f'(x) = \frac{a \cdot a'}{|a|}$.

surface models where the empirical correlations of the points on the reference surface are damped with respect to their geodesic distance. This is achieved by the multiplication of the statistical model kernel with a Gaussian kernel which acts on the surface.

Nonparametric Appearance Model

Since the direct fitting of the shape models to intensity images is difficult, in the second part of this chapter, we presented an approach where the model is fitted to object probabilities. This has the advantage, that systematic appearance structures of an object class can be learned and predicted as foreground. For example the pulp structure inside a wisdom tooth, which barely has meaningful correspondence, can be jointly predicted as object with the dentin and enamel. The object is predicted based on a nonparametric appearance model, where the appearance is learned using random forest regression. Several features, extracted from the image have been used including Harmonic Filters which yield 3D rotation invariant properties.

Incorporate Regions and Boundary

We presented a dissimilarity metric where a ground truth label map can be compared with the target probability map. In addition to the areal object and background part, the boundary of the object is considered as well. This object boundary is used as well for a mesh to probability map metric which we have presented for a plain surface model fitting.

In Chapter 4, we will evaluate the presented methods to show their applicability for medical image segmentation.

4 EXPERIMENTAL RESULTS

This chapter provides qualitative as well as quantitative experimental results for evaluating the presented approaches in this work. The list of content below gives an overview over the structure of this chapter.

4.1 Variational Registration Experiments **68**
4.1.1 POPI Breathing Thorax Model 69
4.1.2 Patellar Surface Registration 75
4.2 Low-rank Gaussian Process Registration vs. Demons Registration **80**
4.3 Wisdom Tooth Shape Model **81**
4.3.1 Shape Model Bias 82
4.4 Wisdom Tooth Segmentation **84**
4.4.1 Evaluation of Segmentation 86
4.4.2 Geodesically Damped Model 91
4.5 Discussion **94**

In the first part Section 4.1, experiments with our variational approach to registration is provided. On the basis of a 4D lung CT dataset, where also landmarks are provided, the approximation performance of the used tensor decomposition techniques are analyzed in detail. The method is compared with the Elastix [49] implementation of the B-spline registration method of Rueckert et al. [80]. For the hybrid landmarks and image registration an extended B-spline method was used which additionally considers the landmark errors as cost term. For a qualitative and visual example, in Section 4.1.2, two femur shapes are registered. Since the bone surface gives insufficient cues about the patellar surface on the femur, landmarks are placed to register it more accurate.

In the second part Section 4.2, the low-rank Gaussian process registration is applied to a femur CT dataset. It is compared to the standard Demons approach [94] and the B-spline registration method [80].

In Section 4.3, we present the statistical tooth model and we also describe its construction. An experiment shows how the model bias can be reduced by combining the statistical model with a generic Gaussian kernel, to reach an overall better expressiveness of the model.

In Section 4.4, the statistical shape model combined with the statistical appearance model is used to segment the wisdom tooth from CBCT images. Our approach is compared to the method of Cremers et al. [29], where a similar appearance model is used. Further, the geodesically damped model is applied to segmentation and compared to the standard empirical surface model.

4.1 VARIATIONAL REGISTRATION EXPERIMENTS

In this section, registration experiments are performed for validating the variational approach for registration. First, a detailed study about the separable filter approximation is provided. We discuss its approximation performance in terms of accuracy and computational aspects. Second, the local filter compression with respect to memory consumption, computational demands as well as approximation accuracy is analyzed. Our method is compared to Elastix [49], where landmarks are incorporated as an additional cost term to the registration functional.

As a quality measurement, the target registration error (TRE) and the dice coefficient (DICE) are used. Further, the singularities in the transformation serve as indicator of how well the deformation field can be inverted. The *singularity count* is determined by counting the voxels where the determinant of the Jacobian is smaller than zero. To compare two deformation fields A and B, the magnitude differences and the vectors directional discrepancy are considered for each vector pair. Following that, we define the accuracy loss:

$$\tau(A,B) = \int_\Omega (\|A(x)\| - \|B(x)\|)^2 + \left\| \frac{A(x)}{\|A(x)\|} \times \frac{B(x)}{\|B(x)\|} \right\|^2 \mathrm{d}x, \quad (4.1)$$

where $\tau(A,A) = 0$ and greater than zero for dissimilar deformation fields.

Since different regularizers and their approximations are compared only, the mean squared differences similarity measure is used for all experiments in this section. Following Thirion [93], gradient descent on $\mathcal{D}$ is performed with the image forces

$$f(\varphi(x)) = -\frac{(I_R \circ \varphi(x) - I_T(x))\nabla I_R \circ \varphi(x)}{\kappa^2 (I_R \circ \varphi(x) - I_T(x))^2 + \|\nabla I_R \circ \varphi(x)\|^2}, \quad (4.2)$$

with κ^2 the reciprocal of the mean squared image spacing. Generally, we set the prior mean function always to a rigid pre-alignment of the images.

ELASTIX CONFIGURATION For the registration with Elastix [49], the B-spline transformation model combined with the mean squares metric and an LBFGS optimizer is used. For the landmark examples the mean squares metric with the "Corresponding Points Euclidean Distance Metric" is applied, which is equivalent to the target registration error.

4.1.1 *POPI Breathing Thorax Model*

In this first experiment, quantitative results by different approximation ranks of the nonseparable exponential kernel ($\alpha = 1$) are shown (*cf.* Section 2.3.1). The filter H_0 has been discretized on a 23^3 voxel mask. The results are compared to the exact method, which is obtained with the same kernel, but without separable filtering. The POPI dataset [99], is used, which contains 10 CT images of a breathing lung. The images have a resolution of $482 \times 360 \times 141$ voxels and a spacing of $0.98 \times 0.98 \times 2\text{mm}^3$. For the experiment, the images have been resampled to $235 \times 175 \times 141$ voxels and scaled to isotropic spacing at 2mm^3. In the experiment, the first image has been chosen to be the reference image. The registrations have been calculated on a single scale level. The experiments have been repeated with increasing rank R of the separable filter approximation from one to four. $R = 1$ corresponds to the rank-one approximation used in Beuthien et al. [14] which serves us as baseline. The exact method corresponds to the algorithm of Long et al. [60] but extended to 3D.

In Figure 26, the image error change during the optimization is illustrated. It has been averaged over the nine registrations. In the first three experiments the convergence rate decreases with increasing rank R, while the resulting image error is getting smaller. One can also observe that for $R \geq 3$ the image error stays nearly the same and is close to the exact method. Moreover, the variance of the image error is getting more narrow with higher R. It can be assumed that for $R > 4$ no significantly improved approximation can be achieved. For a better comparison, all mean curves are again shown together in the last plot.

In Table 1, the results of the experiments are summarized in numerical terms. Evaluating the accuracy loss of Equation 4.1 between the approximations and the exact method, higher rank approximations reach greater accuracy. The CPU time is considerable high for the exact method. With a third of the computational effort, our method achieves a good approximation, without loosing much of accuracy.

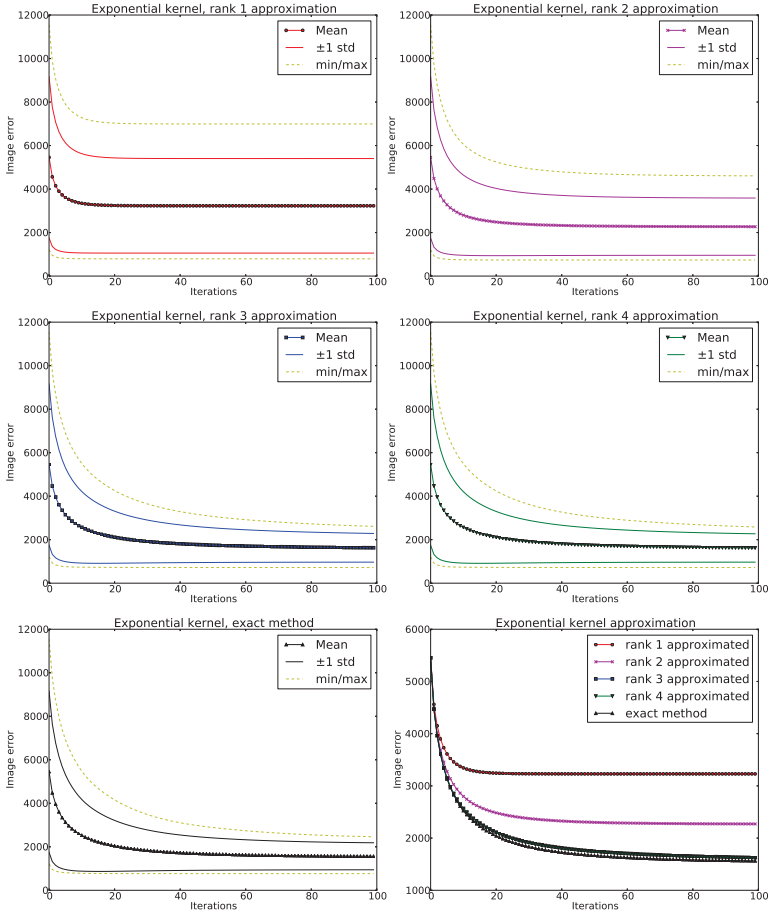


Figure 26: This figure shows the image error averaged over all nine experiments for approximation $R = 1$ to 4 as well as for the exact method. For each experiment the mean error is plotted as well as $\pm$ one standard deviation in a solid style and the max/min as a dashed curve. In the last subfigure the averages for all variants are again shown in one plot.

For a more detailed comparison, the whole experiment is repeated but on different scale levels. The results are listed in Table 2 and Table 3. Note that all quantities are averaged over the nine experiments.

Approx rank	$R = 1$	$R = 2$	$R = 3$	$R = 4$	None
iterations to converge	20	44	68	68	63
final average image error	3588	2524	1808	1799	1737
accuracy loss	1.0033	0.6407	0.2299	0.2250	0.0000
relative CPU time	0.08	0.17	0.25	0.33	1.00

Table 1: This table shows the convergence properties with different approximation ranks. The iterations to convergence in the first row were computed as the number of iterations until the average image error change deceed the value 3. The accuracy loss is calculated by using $\tau(A, B)$ in Equation 4.1 between A the resulting and B the exact method transformation.

	$R = 1$	$R = 2$	$R = 3$	$R = 4$	exact	voxels
CPU time L0 (h)	0.21	0.41	0.61	0.83	3.69	87290
CPU time L1 (h)	1.81	3.54	5.18	7.01	25.71	712530
CPU time L2 (h)	13.88	27.28	40.23	53.19	159.13	5798625

Table 2: This table shows the computational time needed for the different experiments. To show the time complexity depending on the number of voxels in the image, we calculated the experiments on 3 different scale levels L0...L2 and performed 100 iterations per level. The timing is given in average CPU hours needed to perform one registration.

The results show that for nonseparable kernels, a one-rank approximation is not accurate enough to approximate the filter's regularization property. With higher rank, the calculation time increases linear in R. Since the resulting image error as well as the convergence properties using $R = 4$ do not significantly differ from the exact method, we argue that 4 ranks are sufficient to approximate the exponential kernel.

COMPARISON TO ELASTIX For a meaningful comparison to Elastix, three experiments have been performed. First, the smoothness parameter σ of the B-spline transformation model has been tuned to a small TRE ($\sigma = 4$). Second, σ was tuned in order that no singularities (∞)

method	TRE	∞	$\upsilon(A)$	DICE	CPU
rigid alignment	1.673 ± 1.062	0	n/a	0.959	n/a
$k_{\exp}$ $(R = 1)$	1.419 ± 1.059	0	0.094	0.966	15.5
$k_{\exp}$ $(R = 2)$	1.027 ± 0.801	0	0.117	0.974	30.2
$k_{\exp}$ $(R = 3)$	0.588 ± 0.158	0	0.146	0.981	45.9
$k_{\exp}$ $(R = 4)$	0.584 ± 0.149	0	0.148	0.981	59.4
$k_{\exp}$(exact)	0.579 ± 0.103	0	0.198	0.981	188
elastix $(\sigma = 4)$	0.540 ± 0.104	6143.8	0.861	0.987	2.24
elastix $(\sigma = 16)$	0.601 ± 0.150	0	0.398	0.981	2.25
elastix $(\sigma = 64)$	0.819 ± 0.308	0	0.154	0.974	2.23

Table 3: In this table the performance evaluation of the POPI experiment without considering the landmarks are shown. In our method the exponential kernel is used as regularizer. For comparison, in Elastix the B-spline transformation model was used.

are present in the result, but simultaneously for a TRE which is as small as possible ($\sigma = 16$). Finally, the parameter was chosen for a resulting transformation, which is approximately as smooth as the ones obtained by our method ($\sigma = 64$). To quantify the smoothness of a deformation field A the local displacement changes are integrated

$$\upsilon(A) = \int_{\Omega} \int_{\mathfrak{B}_x} \|A(s) - A(x)\| \mathrm{d}s \mathrm{d}x, \tag{4.3}$$

where $\mathfrak{B}_x$ is the neighborhood around x with radius 1. The results in Table 3 show the trade-off in Elastix between TRE and the smoothness of the transformation. σ can be tuned for a small TRE accepting a less smooth transformation or it is chosen such, that the resulting transformation is smooth, but with a higher TRE. However, our method reaches significantly smoother transformations compared to Elastix with a similar TRE. Since we regularize for diffeomorphic transformations it was expected, that compared to Elastix using a small smoothness parameter, no singularities will be present in the results. As soon as σ is increased such, that the transformations are as smooth as in our method, the TRE and DICE performance drops dramatically for Elastix. However, the computational demands for the Elastix experiments are one order of magnitude lower compared to our method.

Ξ	Level	CPU	CPU 1. iter	Mem. (GB)	Masked	τ
without	0	4.05	2.32	10.17	29249	0
	1	9.85	2.63	18.58	54933	0
	2	58.89	5.60	28.61	84424	0
Tucker	0	5.12	2.51	1.11	29249	$6.46 \cdot 10^{-5}$
	1	12.48	3.22	1.44	54933	$2.32 \cdot 10^{-5}$
	2	66.84	6.48	2.55	84424	$9.69 \cdot 10^{-6}$

Table 4: The table shows the resources needed to perform the experiments. For the upper part of the table, no Tucker decomposition of the local filter kernels was performed. In the lower part, the local filters were decomposed using the Tucker decomposition to save memory during caching. H_x was discretized on a 23^3 mask. The approximation ranks were set to $P = Q = R = 5$. The masked voxels (5th column) are the ones, which are in the support of the landmarks. The support limit ξ was set to 10^{-6}.

INCORPORATING LANDMARKS To quantify the efficiency of our filter caching approach, the experiments are performed once more, but including 21 landmarks provided in the POPI dataset. The landmark uncertainty was set to $\sigma = 0.02$. For comparison, the exact method, which combines the separable filtering with $R = 4$ and the landmarks, does no compression. In Table 4, the average resources needed for each experiment are listed. As expected, for the Tucker decomposition, slightly more CPU time is needed. However, it is negligible compared to the memory savings reached with this compression. Furthermore, the approximation of the local filter kernels is nearly perfect resulting in a very small loss of accuracy. The most CPU intensive part in each experiment is the 1st iteration, because initially, all local filter responses have to be calculated. Without the caching scheme therefore, the overall CPU time would explode to CPU weeks.

To compare our hybrid results with Elastix, the hybrid B-spline registration was performed twice, using a small *resp.* a large weight w for the landmark cost term (see Table 5). A large weight results in a smaller TRE while the overall smoothness decreases. Several singularities (indicated as ∞) are present in the Elastix results, while the singularities in our method are negligible. The major advantage of our method becomes apparent with the overall smoothness. Despite the landmark consideration it is much higher than in the Elastix experiment.

method	TRE	∞	$v(A)$	DICE	CPU
rigid alignment	1.673 ± 1.062	0	n/a	0.959	n/a
$k_{\exp}$ (Tucker)	0.484 ± 0.127	0.4	0.144	0.981	80.1
elastix ($\sigma/w = 4/1$)	0.411 ± 0.086	6255.8	0.859	0.987	2.42
elastix ($\sigma/w = 4/64$)	0.024 ± 0.001	2565.9	0.624	0.987	2.41
elastix ($\sigma/w = 16/1$)	0.535 ± 0.089	869.8	0.544	0.984	2.95
elastix ($\sigma/w = 16/64$)	0.159 ± 0.042	948.0	0.534	0.984	2.96
elastix ($\sigma/w = 64/1$)	0.814 ± 0.303	0	0.156	0.974	2.35
elastix ($\sigma/w = 64/64$)	0.796 ± 0.300	0	0.154	0.974	2.34

Table 5: In this table the performance evaluation of the POPI experiment including the landmarks are shown. In our method the exponential kernel is again used as regularizer. In Elastix the B-spline transformation model was used.

Level	speedup (landmarks excl.)	speedup (landmarks incl.)
0	13.8	11.3
1	17.7	12.4
2	17.9	14.9

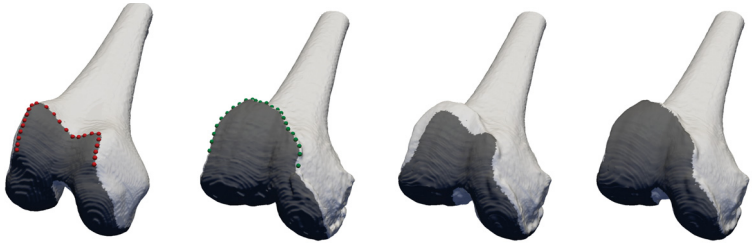
Table 6: The speedups were calculated by (Time (1 process) − Time (24 processes))/Time (24 processes). Therefore, an optimal full parallelizable algorithm would have a speedup of 23, using 24 processes.

DISCRETIZATION ARTIFACTS The TRE could be decreased, but, regarding the small uncertainty on the landmarks a smaller landmark error would have been expected. This discrepancy originates from the discretization of the mean transformation μ_X, which in this experiment leads to a TRE drift of 0.264 ± 0.121.

$$TRE_{\text{drift}} = TRE(\mu_{X\,\text{discretized}}) - TRE(\mu_{X\,\text{exact}}) \tag{4.4}$$

Compared to the experiment in Section 4.1.2, where the resolution is about twice as high the discretization error results in a TRE drift of 0.056 ± 0.002, which is negligible.

A NOTE ON THE PARALLELIZATION Since our method is based on image filtering, it is well suited to perform the filtering for each voxel



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

Figure 27: In this figure, a reference femur shape (a) is registered with the target femur shape (b). The landmarks along the border of the patellar surface are marked in red (reference) and green (target). The patellar surface, marked in dark gray, could not be registered without incorporating the landmarks (c). With the hybrid registration approach presented in this work, such a high-level feature as the patellar surface could be well registered (d).

in parallel. We performed the experiments with 24 processes and reached an average speedup between 15 and 18 as listed in Table 6. Because the landmarks are not uniformly distributed over the image domain, the work load is also not equally distributed to the processes. Therefore, a lower speedup in the hybrid registration experiments has been reached. The actual time needed to perform the experiments is the CPU time listed in the previous tables divided by the speedups listed in Table 6. For example, calculating the 9 registrations on level 2, the exact method took us 2.5 days instead of 8.5 weeks.

4.1.2 *Patellar Surface Registration*

In this section, a 3D experiment is shown where the two femur shapes of Figure 27, are registered. The challenge with this kind of data is that the border of the patellar surface, as described as the gray marked region in Figure 27, is potentially hard to recognize given solely the shape surfaces. Moreover, its variation can be quite large. Therefore, an accurate registration of the patellar surface using automatic algorithms is difficult.

The patellar surfaces of the target and reference bone have been given by experts. By incorporating well-chosen landmarks, our algorithm can

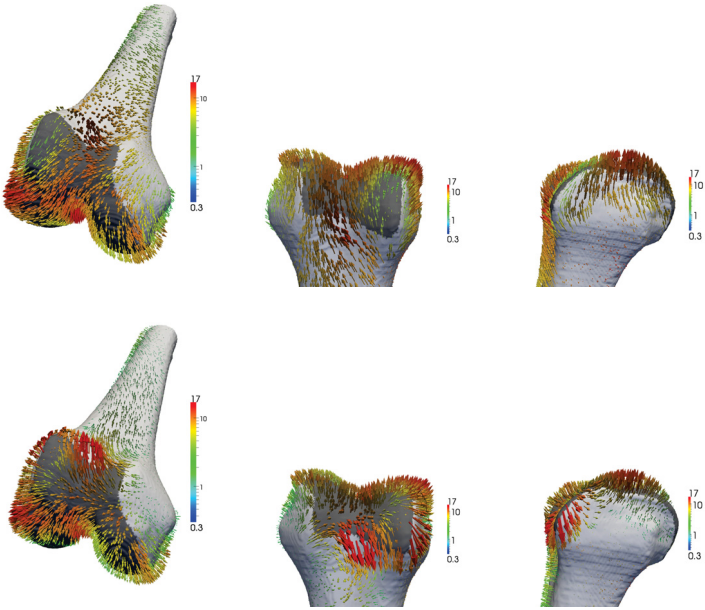


Figure 28: The figures show the displacement field resampled on the reference's surface, depicted as arrows. First row: registration was performed without landmarks. Second row: registration was performed including the landmarks.

be forced to register even the patellar surface correctly. The shapes are represented as signed distance images of $353 \times 327 \times 491$ voxels (isotropic spacing 0.57mm^3 and registered on 5 scale levels l). The scales of the mean transformation μ_X have been chosen to 10^{-l}. For k the Gaussian kernel with $\sigma_G = 1$ have been used and a landmark uncertainty of $\sigma = 0.3 \times 10^{-3}$. We approximated the landmark kernel k_X with $P = Q = R = 5$. For illustration, the experiment is performed once without landmarks and once including the landmarks.

In Figure 28, the displacement fields are shown resampled on the bone surface depicted as arrows. Especially at the upper border of the patellar surface, one can see the strong impact of the landmarks. In Figure 29, the two different displacement fields are again visualized but sampled on the reference landmarks. In Figure 30, the deformed reference shape

(a) Excl. landmarks (b) Incl. landmarks

Figure 29: The figures show the displacement field resampled on the reference's landmarks, depicted as arrows.

method	TRE	∞	$v(A)$	DICE
rigid aligned	12.268 ± 13.627	0	0	0.797
diffeomorphic demons	13.722 ± 31.546	0	0.250	0.989
elastix ($\sigma/w = 1/1$)	**0.005 ± 1.199**	20703	0.314	0.973
our approach	1.225 ± 1.328	**48**	**0.251**	**0.990**

Table 7: Quantitative measures of the femur example.

including the dark gray marked part is plotted. Without considering the landmarks, the border of the patellar surface is clearly misaligned, while it is correctly registered when the landmarks are incorporated.

The same experiment is performed again with Elastix and the summarized results are in Table 7. The parameters of the hybrid B-spline registration were tuned concerning the TRE and DICE performance measures. While Elastix brings the TRE down to nearly zero, a very large amount of singularities (∞) are present in the resulting transformation and the DICE coefficient is rather low. Our method reaches a small TRE as well. Furthermore, the singularity count is very low, the DICE quite high and the deformation field smooth.

SMOOTH MEAN TRANSFORMATION Since multiple reference landmarks can be forced to match one single target landmark by setting the landmark uncertainty σ to zero, μ_X is not guaranteed to be invertible. In Figure 31, an artificial example is shown where a grid is transformed by the mean displacement using different values for σ. Setting σ equal to zero, or too small, results in unfavorable folds and barely

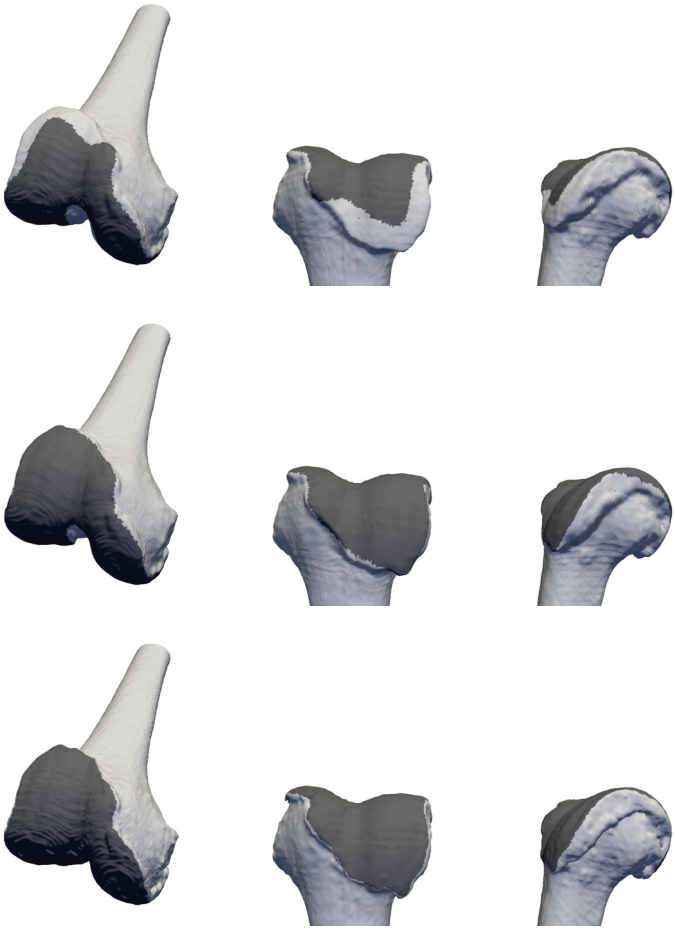


Figure 30: The figures show the deformed reference shape including the colored patellar surface. First row: registration was performed without landmarks. Second row: registration was performed including the landmarks. Third row: ground truth target shape. While there is almost no difference in shape among the rows, the patellar surface in the second row match the one of the ground truth.

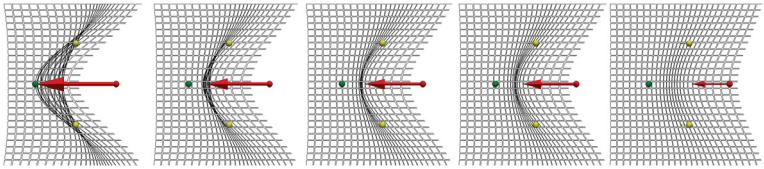


Figure 31: This figure shows a transformed grid (200px^2 and isotropic spacing of 0.1mm^2 with the mean transformation using the Gaussian kernel ($\sigma_G = 6$). There are three landmarks defined as reference and target points (*red* reference, *green* target, the *yellow* ones are equal for both). The uncertainty on the landmarks is increased for the experiments from left to right ($\sigma = 0, \sigma = 0.5 \times 10^{-3}, \sigma = 0.75 \times 10^{-3}, \sigma = 0.1 \times 10^{-2}, \sigma = 0.25 \times 10^{-2}$). The arrows illustrate to which location a point is transformed by the deformation field.

Figure 32: This figure shows the deformed reference femur shape by the mean transformation using the multiscale kernel with five scale levels.

make sense in real wold medical problems. Therefore, in the patellar surface experiment, the parameters have been chosen such that folds in μ_X hardly ever occur. The mean transformed reference shape is shown in Figure 32, where no holes can be identified on the surface.

4.2 LOW-RANK GAUSSIAN PROCESS REGISTRATION VS. DEMONS REGISTRATION

In this experiment, we show the registration performance of the low-rank Gaussian process registration method compared to the standard Demons and the B-spline registration. For this test, CT images of dry femur bones, with a resolution of $176 \times 163 \times 622$ are registered. A reference image was selected which has been registered to 27 target images using the mean squares similarity measure. As a deformation model, a zero mean Gaussian process with a Gaussian kernel ($\sigma_G = 100$) was applied. The low-rank approximation was computed by approximating the first 300 basis functions. The squared image loss was used as well. The registration performance was compared with the ITK implementation of the standard Demons algorithm [46] and the Elastix implementation of th B-spline registration [49]. To have a fair comparison, for each method different parameters have been tested, while the best ones have been used in the comparison. For the Demons method $\sigma = 1.2$ and for the B-Spline registration $\sigma_B = 16$. As performance measure, we use a DICE coefficient, which is computed on manual segmentations of the images. In Figure 34, the reference and three target examples are shown.

Figure 33 shows the results for the different algorithms. The low-rank Gaussian process registration outperforms the Demons algorithm and performs on par with the B-Spline registration. The B-spline registration performs, however, slightly better when a multi-scale strategy is used. This is expected, as the simple Gaussian kernel which is used in the low-rank Gaussian process method only models deformations on a single scale.

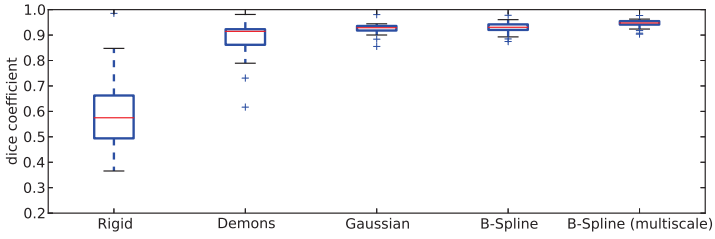


Figure 33: DICE coefficient comparison of femur registration experiments. The rigid alignment was used as baseline.

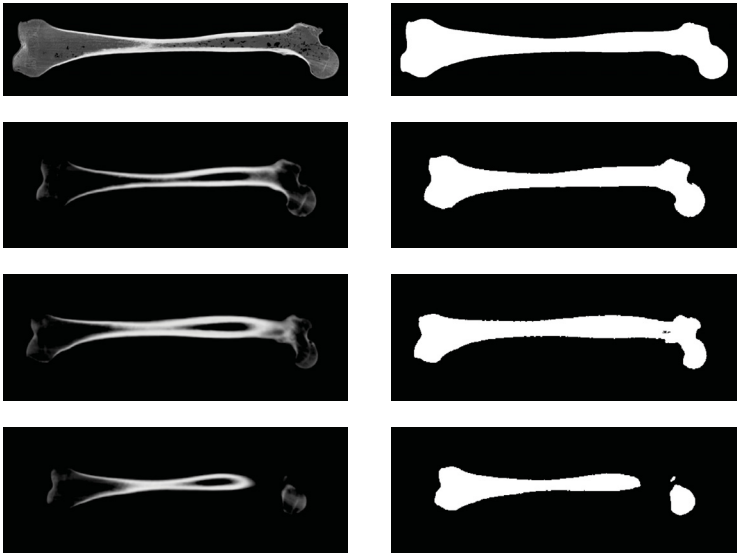


Figure 34: Example femur CT slices. The reference is depicted in the top row. Target examples in the remaining three. In the left column the CT and in the right one the label maps are shown.

4.3 WISDOM TOOTH SHAPE MODEL

The following experiments are based on a dataset of 47 CBCT images of the jaw region with the wisdom tooth as the object of interest. For all images, a ground truth label map is given by experts for following labels: whole tooth, pulp (nerve structure inside the tooth) and the joint structure dentin/enamel (whole tooth except the pulp) including the ones of neighboring teeth. The dataset is restricted to double-rooted teeth. The images have been similarity aligned and their histograms aligned to a training sample. The pulp structure is jointly modeled as tooth shape together with dentin and enamel, because the different number of tracts and hunches makes the registration of the pulp difficult. The point-wise correspondence assumption for this kind of structure is incongruous (*cf.* Figure 21).

STANDARD SHAPE MODEL In the following experiments, different shape models are constructed. The empirical shape models are all built upon a given training set of deformation fields $\mathcal{X} = u_1, \ldots, u_n$. The variation of shape $L_R(x + u_{\boldsymbol{\alpha}}(x))$ is modeled by a linear model

$$u_{\boldsymbol{\alpha}}(x) = \mu(x) + \sum_{i=1}^{n} \alpha_i \phi_i(x), \tag{4.5}$$

where μ is the empirical mean deformation $\mu(x) = \frac{1}{n}\sum_{i=1}^{n} u_i(x)$ and $\phi_i : \mathbb{R}^d \to \mathbb{R}$ the eigenmodes of the training set $\mathcal{X}$ obtained with principal component analysis[1]. With this parametrization over $\boldsymbol{\alpha}$, a simple Gaussian prior is used as regularization $\mathcal{R}[\boldsymbol{\alpha}] = \|\boldsymbol{\alpha}\|^2$.

Having given n training label maps $\{L_i\}_{i=1}^n$, their representation is switched to signed distance maps $\{S_i\}_{i=1}^n$ (*cf.* Section 2.1.3). Using a non-rigid registration method, the shapes are matched to a reference S_R to obtain $\{u_i\}_{i=1}^n$. Which registration method, and how the reference is selected/constructed is mentioned for each specific experiment separately.

4.3.1 *Shape Model Bias*

The aim of the following experiment is to show the impact of the statistical model bias and how it can be reduced to reach better segmentation performance. The distance maps of the 47 given samples have been registered to an arbitrary training sample using the B-spline registration [80] with $\sigma_B = 2$ and 4 resolution levels. An empirical statistical shape model $\mathcal{GP}_{\text{full}}(\mu_{\text{full}}, k_{\text{full}})$ has been constructed using all resulting deformation fields $\{u_i\}_{i=1}^{46}$. Further, the set of deformation fields have been randomly splitted into a training set and a test set with size 19 and 27 respectively. On the basis of this training set, another "reduced" statistical model $\mathcal{GP}_{\text{bias}}(\mu_{\text{bias}}, k_{\text{bias}})$ have been constructed, which simulates a model which does not fully spans the "real" shape variation. Fitting both of the models to the distance maps which correspond to the samples in the test set, the full model clearly outperforms the biased one, which was expected, since the test samples have been included to build the full model (see Figure 35).

Since the biased model is too restrictive, a generic low-rank Gaussian process model $\mathcal{GP}_{\text{generic}}(0, k_G)$ with a Gaussian kernel $\sigma_G = 3$ is constructed. 700 basis functions have been approximated using 7000 uniformly sampled spatial points for the Nyström approximation. Samples

1 The shape models have been built with the open source software Statismo [63].

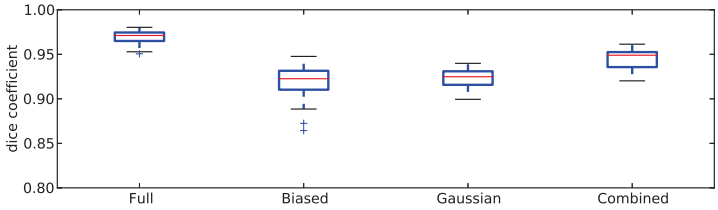
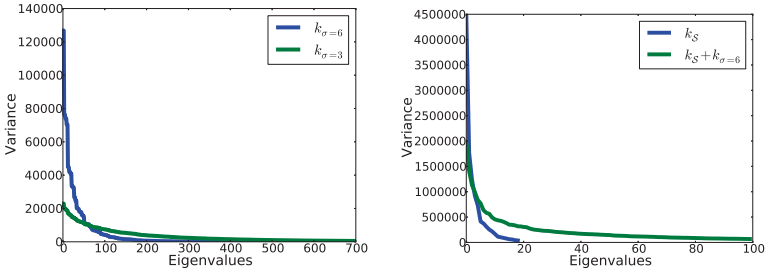


Figure 35: DICE coefficient comparison of different models. For this comparison, no regularization during fitting has been used ($\eta = 0$).



(a) (b)

Figure 36: (a) Eigenvalue spectra of two Gaussian models. The σ have been set to 3 (*blue*) and 6 (*green*) respectively. For both models 700 eigenmodes have been approximated. While in the $\sigma = 6$-model 250 eigenmodes cover 99% of the variance in the $\sigma = 3$-model more than 650 are needed to reach the 99% of the variance. (b) Eigenvalue spectra of the empirical model $k_{\mathcal{S}}$ and a combined model $k_{\mathcal{S}} + k_{\sigma}$ with 100 approximated eigenmodes.

of such a model are visualized in Figure 37. For all experiments the Huber loss function Equation 2.4, has been applied.

Fitting this model to the test set gives similar results than the ones with the biased model *cf.* Figure 35. However, combining the biased and the generic model $\mathcal{GP}_{\text{combined}}(\mu_{\text{bias}}, k_{\text{bias}} + k_G)$ yields a better fitting performance as if they are separately fitted.

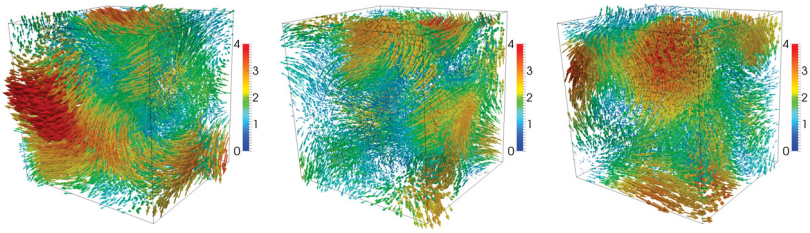


Figure 37: Sample deformation fields of a low-rank Gaussian process model using the Gaussian kernel ($\sigma = 6$mm).

In Figure 36, the eigenvalue spectra of different generic models and a bias reduced model is plotted. Restricting the admissible functions to be smoother (using a larger σ of the Gaussian kernel) the spectrum decreases faster. In the bias reduced model, one can observe, that the overall variance have increased compared to the empirical model.

4.4 WISDOM TOOTH SEGMENTATION

In this section, the segmentation performance using the statistical tooth shape model and the statistical appearance model presented in this work is evaluated. The performance of three different object prediction strategies are evaluated. Our approach is compared to the classical Active Shape Model of Cootes et al. [25] which serves as baseline[2]. Moreover, we show comparisons with the method of Cremers et al. [29] and an extended version of their approach using the object prediction presented in this work. The results are evaluated using two quantitative measures. For the first measure, the boundary of the ground truth labels as well as the ones of the segmentation results are sampled to get meshes, which can be compared. As a quantitative measure, we apply the Bidirectional Local Distance (BLD) which was proposed by Kim et al. [48] in order to compare segmentations. Additionally, the segmentation performance is evaluated using the DICE coefficient between the ground truth and the segmentation label maps.

To build the statistical shape model, 26 training samples are registered to an arbitrary training sample using B-spline registration of the signed distance maps. Similar to the atlas construction of Frangi et al. [33],

2 We used the implementation of Kroon [53].

the average deformation is applied to the reference label, resulting in an average label map, which again is used as final reference for registration of all training samples. In this way, the model bias towards the reference is reduced.

OBJECT PREDICTION Our method is tested with two object prediction setups:

INTENSITY A random forest classifier[3] is trained on the whole tooth labels using the intensity values only.

IMPROVED Two random forests are separately trained; one on the pulp and one on the dentin/enamel labels. The following features have been used:

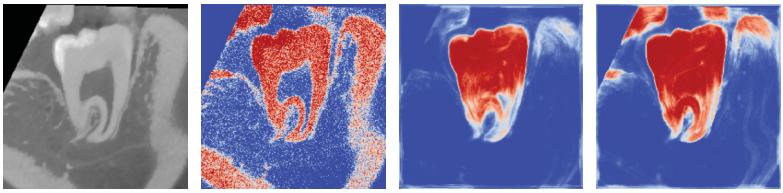
- intensity values
- Laplacian of Gaussian on three scales ($\sigma = \{0.5, 1, 2\}$)
- Gaussian gradient magnitudes
- Sobel edges
- Multi-channel Harmonic Filters features [82] of the order 3 on four scales ($\sigma = \{2, 5, 8, 11\}, \gamma = \{1, 1, 1, 1\}$)

For prediction, the maximum of the two classifier outputs is used.

All random forests have been trained with 50,000 points per label, uniformly sampled as training data (a border of 0.75mm has been excluded from sampling to reduce image boundary effects). 256 trees have been learned per forest, where 30% of the sampled points have been used per tree. The Gini index is used as criterion for the splitting function (more details can be found in [51]).

In Figure 38, example probability maps of the two different prediction methods *intensity* and *improved* applied to a test sample, are shown. Additionally, another probability map is compared where the prediction is based on the same features, which have been used in *improved*, but with only one single random forest trained on the whole tooth label. In the *intensity* probability map, the pulp could not be predicted as belonging to the object, because it shares intensity values with the background. Furthermore, the bone structure of the jaw is not fully predicted as background. Overall, the noise from the original image is passed over to the probability map, which originates in the simple intensity features. On the tooth probability map, the background and even the neighboring teeth (upper left of the tooth crown) could be greatly reduced. But,

3 We used the random forest implementation of the computer vision library VIGRA [51].



(a) I_T (b) $P_{\text{Intensity}}$ (c) P_{Tooth} (d) P_{Improved}

Figure 38: Example probability maps. (a) denoised CBCT image, (b) object probabilities based on *intensity* features, (c) object probabilities based on *improved* features but with only one random forest and (d) object probabilities based on *improved* prediction. Although the background is better predicted in (c) there are uncertainties left in the region of the roots. In (d), the tooth is predicted more accurate, while the neighboring teeth are predicted as object as well.

there is some uncertainty left in the region of the roots. In the *improved* probability map, the background is predicted accurately, while the neighboring teeth are predicted as object as well. Nevertheless, the main objective, namely the overall tooth structure including the roots is better predicted than in the other two probability maps.

In Table 8, we quantitatively evaluate the object prediction improvement obtained with the presented methods. We applied two different performance measures: Υ_{obj}, which considers the object prediction only and $\Upsilon_{\text{obj}}\Upsilon_{\text{bg}}$, which also incorporates the background prediction, where:

$$\Upsilon_{\text{obj}} = \frac{\sum_x L(x)P(x)}{\sum_x L(x)}, \quad \Upsilon_{\text{bg}} = \frac{\sum_x (1-L(x))(1-P(x))}{\sum_x (1-L(x))}. \tag{4.6}$$

The results show, that the feature based methods greatly improve the prediction. The *improved* features improve mainly the object prediction accepting little less accurate background prediction.

4.4.1 *Evaluation of Segmentation*

In Table 9, our approach is quantitatively evaluated and compared to the Active Shape Model approach [25]. Further, it is compared to the method of Cremers et al. [29] where only the model fitting without the

	Method	Test	Training
Υ_{obj}	$P_{\mathrm{Intensity}}$	$0.658 \pm 8.19e-3$	$0.692 \pm 9.41e-3$
	P_{Tooth}	$0.810 \pm 8.77e-3$	$0.879 \pm 2.78e-3$
	P_{Improved}	$0.834 \pm 5.77e-3$	$0.886 \pm 4.71e-3$
$\Upsilon_{\mathrm{obj}}\Upsilon_{\mathrm{bg}}$	$P_{\mathrm{Intensity}}$	$0.601 \pm 1.44e-3$	$0.605 \pm 1.81e-3$
	P_{Tooth}	$0.776 \pm 6.75e-3$	$0.842 \pm 1.49e-3$
	P_{Improved}	$0.794 \pm 2.31e-3$	$0.825 \pm 1.11e-3$

Table 8: Numerical results: average performance ± 1 standard deviation.

global optimization was used, since the images are pre-aligned. For our method, the parameters have been chosen as follows: weighting of the regularization $\gamma = 0.0001$ and the trade-off between region and boundary terms $\lambda = 0.75$. Image gradients have been calculated by convolving the first order derivative of a Gaussian, where $\sigma = 0.15$.

In the upper part of Table 9, the BLD between the ground truth and resulting shapes was used as performance measure, while in the bottom part, the DICE coefficient on the whole image domain serves as segmentation accuracy measure. The ASM method was performed on the histogram matched images and the method of Cremers et al. [29] using the *intensity* and the *improved* object prediction. Two additional variants for our approach are evaluated: one using an arbitrary training label map as reference (*arb. ref.*) and one without the boundary term ($\lambda = 1$) (*no bound.*). The two last rows show the maximum capability of the segmentation methods where the ground truth label maps are used as probability maps (*ground truth*).

On average, the *improved* object prediction could always improve the test segmentation performance, compared to the *intensity* experiments, even in the method of Cremers et al. [29]. The two additional experiments *arb. ref.* and *no bound* show that the constructed reference shape as well as the incorporation of the boundary term improve the test segmentation accuracy. Since the ASM method is based on the object boundary and only considers the image intensities, where the boundaries are barely pronounced, its segmentation performance is minor. Compared to the intensity model of Cremers et al. [29] the deformation model and the incorporation of the object boundary in our method leads to better segmentation results.

Method	Test (BLD)	Training (BLD)
ASM	$0.365/0.418 \pm 18.07e-3$	$0.340/0.376 \pm 15.72e-3$
Cremers *et al.* (intensity)	$0.354/0.366 \pm 3.996e-3$	$0.349/0.345 \pm 5.530e-3$
Cremers *et al.* (improved)	$0.322/0.341 \pm 3.720e-3$	$0.337/0.328 \pm 5.188e-3$
Our (intensity, arb. ref.)	$0.272/0.387 \pm 65.48e-3$	$0.167/0.189 \pm 3.174e-3$
Our (intensity, no bound.)	$0.266/0.365 \pm 44.02e-3$	$0.156/0.170 \pm 2.463e-3$
Our (intensity)	$0.268/0.350 \pm 38.61e-3$	$0.159/0.174 \pm 2.134e-3$
Our (improved, arb. ref.)	$0.306/0.362 \pm 20.29e-3$	$0.197/0.235 \pm 7.487e-3$
Our (improved, no bound.)	$0.263/0.338 \pm 22.47e-3$	$0.184/0.193 \pm 1.168e-3$
Our (improved)	$0.262/0.334 \pm 20.75e-3$	$0.186/0.199 \pm 1.453e-3$
Cremers *et al.* (ground truth)	$0.293/0.310 \pm 3.987e-3$	$0.242/0.233 \pm 5.717e-3$
Our (ground truth)	$0.192/0.211 \pm 3.276e-3$	$0.090/0.091 \pm 0.131e-3$

Method	Test (DICE)	Training (DICE)
Cremers *et al.* (intensity)	$0.883/0.881 \pm 0.397e-3$	$0.887/0.886 \pm 0.683e-3$
Cremers *et al.* (improved)	$0.892/0.892 \pm 0.260e-3$	$0.891/0.896 \pm 0.435e-3$
Our (intensity, arb. ref.)	$0.919/0.905 \pm 1.538e-3$	$0.946/0.942 \pm 0.256e-3$
Our (intensity, no bound.)	$0.920/0.907 \pm 1.067e-3$	$0.950/0.946 \pm 0.193e-3$
Our (intensity)	$0.920/0.910 \pm 0.860e-3$	$0.949/0.945 \pm 0.165e-3$
Our (improved, arb. ref.)	$0.912/0.904 \pm 0.511e-3$	$0.937/0.932 \pm 0.302e-3$
Our (improved, no bound.)	$0.919/0.909 \pm 0.638e-3$	$0.941/0.941 \pm 0.063e-3$
Our (improved)	$0.918/0.910 \pm 0.542e-3$	$0.940/0.939 \pm 0.068e-3$
Cremers *et al.* (ground truth)	$0.901/0.901 \pm 0.278e-3$	$0.925/0.927 \pm 0.536e-3$
Our (ground truth)	$0.939/0.936 \pm 0.155e-3$	$0.968/0.968 \pm 0.015e-3$

Table 9: Numerical results: median / average performance measure $\pm$ 1 standard deviation.

In Figure 39 and Figure 40, qualitative results are shown. Three difficult test examples which have been well segmented by our method are examined in more detail. The first is difficult, because of touching root tips. In the second one, there is almost no contrast between jaw and root structure. And in the last example there is a lot of noise present in the region of the roots. The last column shows the ground truth shape (blue) overlayed with our segmented shape (yellow).

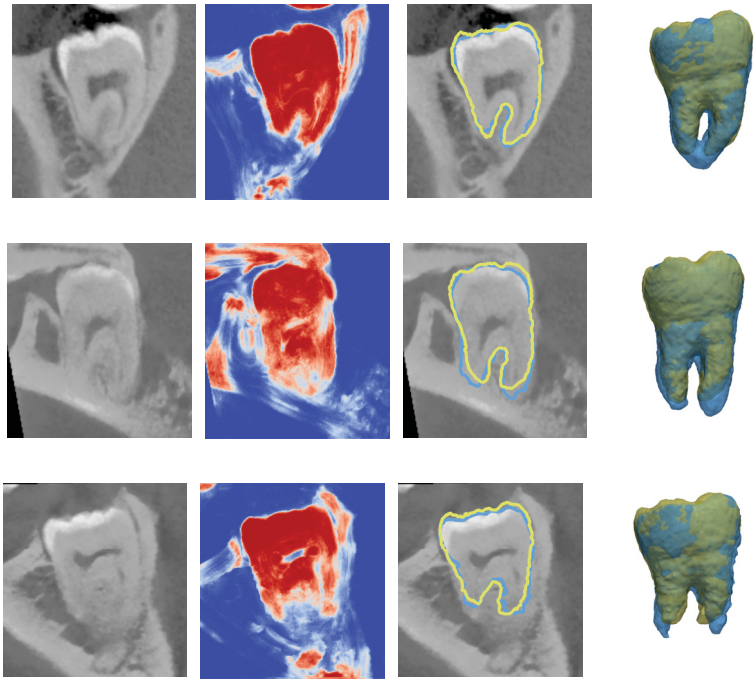


Figure 39: Three difficult test images. In the first column the original CT images are shown. The second column shows the *improved* probability maps. The third column shows the ground truth label in *light blue* and the segmentation result in *yellow*. In the last column the teeth are visualized as 3D meshes in the same colors.

INCORPORATING LANDMARKS The variance in the shape comparison shows that there are some outliers in our segmentation results. In Figure 41, three bad examples are depicted. In such cases, it is important to have the possibility to refine the result given some landmarks. For these three examples the following additional experiments have been performed. The empirical model k_S, built from the 26 training samples, was conditioned on landmarks using the conditional model of Section 2.4.3. Further, another conditional model based on the combined model $k_S + k_G$ was constructed. The results are visualized in Figure 42

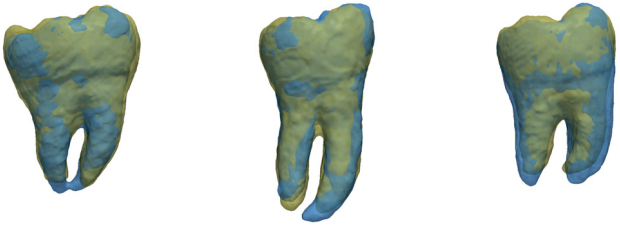
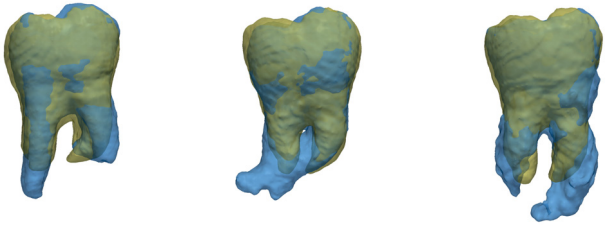


Figure 40: Three examples where the segmentation succeeded well. The ground truth and the segmented shapes are colorized in *blue* and *yellow* respectively.



(a) Sample 1 (b) Sample 2 (c) Sample 3

Figure 41: Three examples where the segmentation is bad. The ground truth and the segmented shapes are colorized in *blue* and *yellow* respectively.

and summarized in Table 10, where the resulting DICE coefficients are listed when fitting the conditional models to the three target samples using the *improved* probability map.

Despite the combined model has more flexibility (*cf.* the experiment in Section 4.3.1), in the this setting, it is too flexible and results in worse DICEs than the simple empirical model. However, if landmarks are placed the constrained combined model results in significantly better segmentations. For the Sample 1, 5 landmarks were enough. For the other two samples, 9 landmarks have been placed. For all three samples,

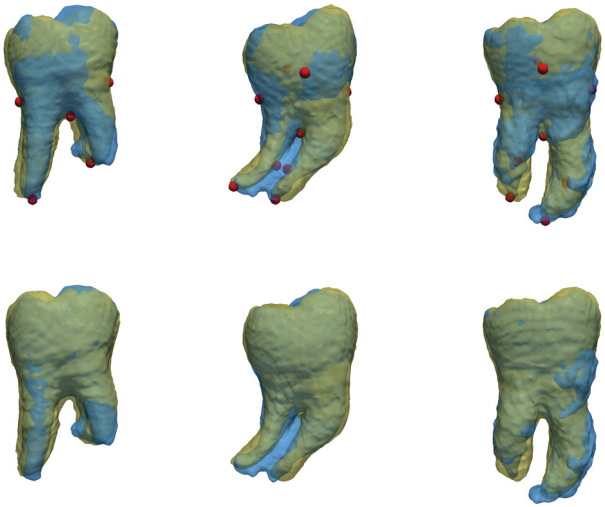


Figure 42: In the columns again the three outliers are depicted. In the first row, the mean of the constrained combined model is shown, where the landmarks are visualized as *red* dots. In the second row, the constrained combined model was fitted to the *improved* probability map. The ground truth and the segmented shapes are colorized in *blue* and *yellow* respectively.

the fitting with the constrained combined model could not improve the segmentations and look very similar compared with their mean.

4.4.2 *Geodesically Damped Model*

In this last experiment, a bias reduced statistical surface model has been built, where the point-correlations are damped with respect to their geodesic distance on the surface. With this experiment, the flexibility of the low-rank Gaussian process framework is shown, since a *complex* kernel which acts on the surface is approximated.

The same 26 training samples as in the previous experiment have been used to construct the statistical shape model, which have been used in the wisdom tooth segmentation experiment in the previous section. They

Method	Sample 1	Sample 2	Sample 3
Empirical (fitted)	0.896908	0.896476	0.866706
Constrained (mean)	0.877183	0.818361	0.885599
Constrained (fitted)	0.901916	0.849155	0.889986
Combined (fitted)	0.878206	0.842832	0.755303
Constrained combined (mean)	0.905973	0.874718	0.898643
Constrained combined (fitted)	0.905468	0.870980	0.857792

Table 10: Constrained: conditional model $k_{\mathcal{X}}$ with an uncertainty on the landmarks of $\sigma = 0.1$. Combined: combined model $k_{\text{combined}} = k_{\mathcal{S}} + k_G$ with a Gaussian $\sigma_G = 3$. Constrained combined: combined model k_{combined} conditioned on the landmarks.

have been registered to one arbitrary training sample Γ_R using B-spline registration. The empirical shape model, which is only defined on the surface domain Γ_R is denoted as $k_{\mathcal{S}}$, which is discretized on 28K points. The geodesic kernel k_Γ on the reference surface Γ_R and $\sigma = 3$ has been approximated with 100 basis functions and 3000 points for the Nyström method. Similarly, the combined kernel $k_{\mathcal{S}} \cdot k_\Gamma$ has been approximated with the same parameters. As a comparison, the same models but with a Gaussian kernel k_G and $\sigma = 3$ have been built.

In this experiment, we mainly show the different behaviors of the geodesic *resp.* Gaussian model and thus, we have not fine tuned the parameters to the best possible segmentation results. For the model fitting, the similarity metric of Equation 3.23 which only incorporates the boundary was used. A first order gradient descent optimizer was used to minimize the objective functional

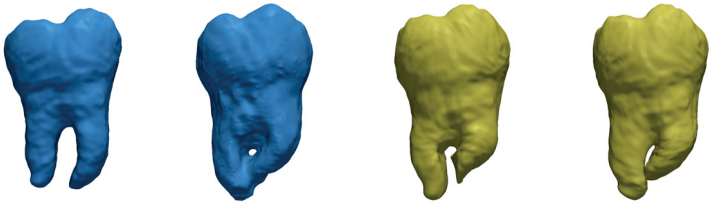
$$\arg\min_{\boldsymbol{\alpha}} \mathcal{D}_\Gamma[\boldsymbol{\alpha}] + \lambda \|\boldsymbol{\alpha}\|^2, \tag{4.7}$$

where λ was set to $0.1e{-}4$ and the σ for the Gaussian gradient magnitude of the object probability map to 0.3. In Table 11, the bidirectional local distance (BLD)[48] is evaluated for the different results, where "(comb)" indicates the combined kernel. To show the maximum capability of the methods, in the bottom part of the table, the ground truth label maps have been used as object probability maps.

Compared to the empirical model, all others could improve the segmentation performance. Especially the combined models greatly improve the segmentation. On average, the geodesic kernel performs on par with

	Model	Test (BLD)	Training (BLD)
Probability Map	Empirical	$0.288/0.390 \pm 71.5e-3$	$0.160/0.179 \pm 1.60e-3$
	Geodesic	$0.256/0.391 \pm 121e-3$	$0.214/0.231 \pm 5.39e-3$
	Geodesic (comb)	$0.240/0.301 \pm 35.4e-3$	$0.179/0.187 \pm 2.53e-3$
	Gaussian	$0.269/0.337 \pm 34.7e-3$	$0.246/0.268 \pm 10.3e-3$
	Gaussian (comb)	$0.251/0.301 \pm 27.1e-3$	$0.201/0.209 \pm 1.88e-3$
Ground Truth	Empirical	$0.223/0.278 \pm 24.2e-3$	$0.139/0.144 \pm 0.54e-3$
	Geodesic	$0.133/0.153 \pm 2.71e-3$	$0.135/0.157 \pm 4.41e-3$
	Geodesic (comb)	$0.136/0.165 \pm 5.42e-3$	$0.115/0.117 \pm 0.19e-3$
	Gaussian	$0.130/0.142 \pm 1.11e-3$	$0.131/0.148 \pm 2.74e-3$
	Gaussian (comb)	$0.150/0.175 \pm 5.02e-3$	$0.143/0.152 \pm 1.99e-3$

Table 11: Numerical results: median / average performance ± 1 standard deviation.



(a) Reference (b) Target (c) k_G (d) k_Γ

Figure 43: (a), (b) depict the reference and the target surface. (c) is the registration where the Gaussian kernel has been used. (d) is the registration where the geodesic kernel has been used.

the Gaussian kernel. We picked an example of the ground truth experiments where either only the geodesic or Gaussian kernel have been used. The results are shown in Figure 43. While in the Gaussian case the roots cannot be "pressed" together, in the geodesic case they can, because topologically, the root tips are treated as being wide apart and thus uncorrelated.

4.5 DISCUSSION

In the first part of this chapter, we evaluated the variational registration approach where nonseparable and nonstationary kernels are applied. On the basis of the POPI breathing thorax dataset, we have shown the improvement of the approximation performance when a higher rank approximation of the filter kernels is used. We compared our method to the B-spline registration of Elastix, where we have shown the tradeoff between smoothness and accuracy of the resulting registration. The same experiments have been performed including landmark displacements. Compared to Elastix with a similar target registration error, we obtain smoother deformation fields accepting more computational power of course. A qualitative patellar surface registration example is shown where similar registration properties can be observed. Accepting a small TRE, our method yields smoother and more accurate registration results compared to Elastix.

We further compared the Demons registration to the low-rank Gaussian process registration with a Gaussian kernel, applying them to a dry femur dataset. Besides the better performance over the Demons registration our approach performs on par with the B-spline registration. The flexibility of our approach has been shown with the bias reduced statistical shape model, where we have shown that the composition of the statistical model and a generic Gaussian model yields better fitting performance.

In the second part of this chapter, we have evaluated the object prediction and the segmentation performance of our method. We showed prediction improvement using neighborhood dependent features. Our model fitting outperforms the classical Active Shape Model approach and the method of Cremers et al. [29]. We have shown, that the bootstrap approach to construct a reference and the incorporation of the object boundary into the minimization function, further improves the segmentation performance. Finally, we used landmark displacements and the bias reduced model to improve the segmentation performance of outliers.

In the last experiment, we evaluated and compared different surface models. The generic and the combined models outperform the empirical statistical surface model. In a case study of a difficult tooth shape, where the roots touch each other, we have shown that the geodesically damped model is expressive enough to “press” the roots together.

OUTLOOK There is room for improvements for the joint approach to object prediction and model fitting. Since our method for object prediction is not optimized at all, we think a clever feature selection in the random forest learning, as it is done by Yaqub et al. [104], would significantly accelerate the object prediction. Further, a grid search to optimize the tree-count as well as the tree depth in a forest would relax the computational demands as well.

Concerning the model fitting, we are also interested to train and use the random forests to drive the shape evolution as it is done in Cootes et al. [24]. Also interesting would be to learn a forest for each point on the surface and project the predicted surface into the statistical model, *cf.* Lindner et al. [59]. This approach could be combined with Margeta et al. [65], where iteratively after each fit new random forests are learned for the prediction.

5

CONCLUSIONS

In this thesis, we examined the statistical model fitting approach to wisdom tooth segmentation from CBCT images. For modeling the variation of shape, a main part was the registration. We presented and discussed two major approaches. First, the nonparametric variational approach, where a partial differential equation is solved, whose solution gives rise to the registration. Using tensor decomposition techniques, smoothness priors which are nonseparable and nonstationary could be efficiently implemented, allowing us to additionally incorporate prior landmark displacements. Especially, when allowing strong generic variations in the transformation, the variational approach, equipped with the diffeomorphic constraint, is efficient to compute and yields smooth registration results. However, if filter kernels have a long range support and are highly spatially variable, the computational costs grow considerably.

A second approach to registration has been presented where the registration problem is directly solved in the basis of a reproducing kernel. Using low-rank approximation and subsampling of the kernel basis, an efficient and flexible algorithm for registration is derived. Thus, each sufficiently smooth kernel function can be approximated and also conditioned on landmark displacements. In particular, the same algorithm can be used for registration, where a generic smoothness prior is applied, as for statistical model fitting. The approach is however limited to kernels, whose eigenvalue spectrum decreases rapidly, *i.e.* kernels which favor small variations in the deformations. Otherwise, a large number of basis functions have to be approximated which becomes computational demanding.

We presented two different approaches to relax the statistical shape model bias. With the help of the introduced low-rank Gaussian process approach, we combined the empirical shape model and a generic Gaussian deformation prior and showed the gain of flexibility using this model. In addition, we built a statistical surface model, where the point correlations are damped with respect to their geodesic distance on the surface. In the segmentation experiments, the combined models yield better segmentation performance compared to the empirical and generic deformation priors. Manipulating the statistical covariance becomes feasible

because of the presented low-rank method. Otherwise, the full covariance matrix would have to be explicitly calculated.

For fitting the model to unseen data, we presented a nonparametric appearance model based on random forest regression. Instead of using the intensity images, we fit the models to object probability maps. With the help of the appearance model, a guess about the object is provided for each voxel separately. As such, the fit of the statistical model yields a global consistency across the prior guesses. In the experiments, we showed significant segmentation performance gain using this combination of appearance and shape model.

5.1 FUTURE WORK

In the presented variational registration approach, an efficient nonstationary filtering approach is derived to solving a partial differential equation using the *landmark* kernel as smoothness prior. Since the initial filter kernel construction is still computational demanding it would be interesting to investigate in the approach of Cahill et al. [20]. They presented a coupled PDE system to locally adapt the regularization strength. A weighting function would have to be built which implements the landmark-aware part of the *landmark* kernel.

In our hybrid landmark and image registration approach, the landmark regression is not smoothly invertible. An additional extension to this approach concerning diffeomorphic image registration would be to add the diffeomorphic constraint to the landmark mean function. One possibility would be the Gaussian process regression with fluid hyperpriors of Girdziušas and Laaksonen [37].

In the RKHS framework for registration, we considered kernels which are stationary. However, it would be interesting to use generic kernel models which adapt their regularization strength to the image location. In [42, 75], spatially aware kernels have been presented where the basis functions have been locally weighted using spectral tempering. Schmah et al. [83] have even extended the diffeomorphic registration framework LDDMM to spatially varying kernels. To integrate the low-rank approximation into their framework would be very interesting.

A detailed analysis of the Nyström approximation performance concerning our registration approach would be interesting. For example, an intuition about how to choose the number of sampled points in the Nyström approximation given a kernel function would be helpful.

Our appearance model is integrated in the segmentation as a preprocessing step. It would be interesting to intertwine it more with the shape

model fitting. Margeta et al. [65] presented an approach where the intensity normalized images are first rigidly initialized. Subsequently, again a random forest is trained. Another interesting idea presented by Lindner et al. [59] is to directly predict the shape surface using random forest regression voting. In a similar approach by Cootes et al. [24], the random forest regression voting is used for the shape evolution in the fitting.

Concerning the appearance model, additional experiments for selecting optimal features and a grid-search for estimating the best tree depth *resp.* tree count in the random forest would be required to optimize the prediction performance. Additionally, it would be interesting how sensitive to the initial alignment the prediction is.

Part III

APPENDIX

A

KERNEL REGISTRATION FRAMEWORK

A.1 TRANSLATION-INVARIANT KERNEL FUNCTIONS

To construct a discrete filter kernel, the kernel function has to be normalized in a way that the integral over the whole domain of the kernel function is 1. Let k be our kernel function, the following condition has to be fulfilled:

$$\frac{1}{c} \cdot \int_{\mathbb{R}^d} k\left(0, \mathbf{x}\right) \mathrm{d}\mathbf{x} = 1. \tag{A.1}$$

Therefore c has to be determined for each kernel function.

GAUSSIAN KERNEL

$$k\left(\mathbf{x}, \mathbf{y}\right) = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^d} \cdot \exp\left(\frac{-\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}\right), \tag{A.2}$$

where d is the number of dimensions.

EXPONENTIAL KERNEL

For the c corresponding to the exponential kernel, the following integral has to be calculated:

$$c = \int_{\mathbb{R}^d} \exp\left(\frac{-\|\mathbf{x}-\mathbf{y}\|}{\sigma}\right) \mathrm{d}\mathbf{x}. \tag{A.3}$$

An analytical solution can be derived by a transformation into spherical coordinates to find a stamm function for:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left(\frac{-\sqrt{x^2+y^2+z^2}}{\sigma}\right) \mathrm{dxdydz}$$

in the 3 dimensional case. The transform $(x, y, z) \rightarrow (r, \theta, \varphi)$ is:

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} \sin\theta\cos\varphi \\ \sin\theta\sin\varphi \\ \cos\theta \end{bmatrix} \cdot r.$$

The volume element then is:

$$\mathrm{dV} = \mathrm{dxdydz} = r^2 \sin\theta \mathrm{d}\theta\mathrm{d}\varphi\mathrm{d}r.$$

The transformed integral of Equation A.3 becomes:

$$\int_{r=0}^{\infty}\int_{\theta=0}^{\pi}\int_{\varphi=0}^{2\pi} e^{\frac{-r}{\sigma}} \cdot r^2 \sin\theta \mathrm{d}\varphi\mathrm{d}\theta\mathrm{d}r. \tag{A.4}$$

The solution to Equation A.3 therefore is (for 3D):

$$\underbrace{\int_{\theta}\int_{\varphi} \sin\theta \mathrm{d}\theta\mathrm{d}\varphi}_{4\pi} \int_{0}^{\infty} e^{\frac{-r}{\sigma}} r^2 \mathrm{d}r = 4\pi \cdot 2\sigma^3$$

and for 2D:

$$\mathrm{dV} = \mathrm{dxdy} = r^2 \mathrm{d}r\mathrm{d}\varphi$$

$$\underbrace{\int_{\varphi=0}^{2\pi} 1 \cdot}_{2\pi} \int_{0}^{\infty} e^{\frac{-r}{\sigma}} r\mathrm{d}r = 2\pi\sigma^2.$$

Finally the normalized kernel function is (in 3D):

$$k(\mathbf{x},\mathbf{y}) = \frac{1}{4\pi \cdot 2\sigma^3} \cdot \exp\left(\frac{-\|\mathbf{x}-\mathbf{y}\|}{\sigma}\right) \tag{A.5}$$

(and in 2D)

$$k(\mathbf{x},\mathbf{y}) = \frac{1}{2\pi\sigma^2} \cdot \exp\left(\frac{-\|\mathbf{x}-\mathbf{y}\|}{\sigma}\right) \tag{A.6}$$

A.2 KERNEL COMPOSITIONS

Given valid kernels $k_1(x,y)$ and $k_2(x,y)$, the following new kernels will also be valid (*cf.* Bishop et al. [15]):

$$k(x,y) = ck_1(x,y) \tag{A.7}$$
$$k(x,y) = f(x)k_1(x,y)f(y) \tag{A.8}$$
$$k(x,y) = q(k_1(x,y)) \tag{A.9}$$
$$k(x,y) = \exp(k_1(x,y)) \tag{A.10}$$
$$k(x,y) = k_1(x,y) + k_2(x,y) \tag{A.11}$$
$$k(x,y) = k_1(x,y)k_2(x,y) \tag{A.12}$$
$$k(x,y) = k_3(\phi(x),\phi(y)) \tag{A.13}$$
$$k(x,y) = x^T\mathbf{A}x \tag{A.14}$$

for $c \in \mathbb{R}^+$ as a constant, f is any function, q is a polynomial with nonnegative coefficients, ϕ is a function from x to $\mathbb{R}^M$, k_3 is a valid kernel in $\mathbb{R}^M$ and $\mathbf{A}$ is a symmetric positive semidefinite matrix.

A.3 MINIMIZATION IN RKHS REGISTRATION

In general the objective

$$\arg\min_{u \in \mathcal{H}_k} \frac{1}{2N} \sum_{i=1}^{N} (I_T(x_i + u(x_i)) - I_R(x_i))^2 + \frac{\lambda}{2} \|u\|_k^2 \tag{A.15}$$

can be minimized by gradient descent optimization methods, where the gradient is given by

$$\frac{1}{N} \sum_{i=1}^{N} (I_T(x_i+u(x_i))-I_R(x_i)) \nabla I_T(x_i+u(x_i))-I_R(x_i)) \mathbf{J}_u(x_i) + \frac{\lambda}{2} \frac{\partial}{\partial \boldsymbol{\alpha}} \|u\|_k^2. \tag{A.16}$$

Recall,

$$u(x) = \mu(x) + \sum_{i=1}^{N} \boldsymbol{\alpha}^T K_x. \tag{A.17}$$

Thus, the derivative of this RKHS transform with respect to $\boldsymbol{\alpha}$ is

$$\frac{\partial}{\partial \boldsymbol{\alpha}} u(x) = K_x, \tag{A.18}$$

and of the RKHS norm

$$\frac{\partial}{\partial \boldsymbol{\alpha}} \|u\|_k^2 = 2\boldsymbol{\alpha}^T K_x. \tag{A.19}$$

A.4 PCA AND SVD

Usually, the number of points N which need to be sampled is much larger than the number of examples p ($N >> p$). Each matrix $\mathbf{S} \in \mathbb{R}^{Nd \times p}$, $Nd \geq p$ can be decomposed into

$$\mathbf{S} = \mathbf{U}_\mathcal{S} \mathbf{D}_\mathcal{S} \mathbf{V}_\mathcal{S}^T, \tag{A.20}$$

where $\mathbf{U}_\mathcal{S} \in \mathbb{R}^{Nd \times p}$ is an orthonormal matrix, $\mathbf{D}_\mathcal{S}$ is diagonal and $\mathbf{V}_\mathcal{S} \in \mathbb{R}^{p \times p}$ is orthonormal as well. Thus, $\frac{1}{p}\mathbf{S}\mathbf{S}^T$ can be reconstructed with

$$\frac{1}{p}\mathbf{S}\mathbf{S}^T = \frac{1}{p}\mathbf{U}_\mathcal{S}\mathbf{D}_\mathcal{S}\mathbf{V}_\mathcal{S}^T\mathbf{V}_\mathcal{S}\mathbf{D}_\mathcal{S}\mathbf{U}_\mathcal{S}^T = \frac{1}{p}\mathbf{U}_\mathcal{S}\mathbf{D}_\mathcal{S}^2\mathbf{U}_\mathcal{S}^T. \tag{A.21}$$

Since the decomposition in Equation A.20 can be computed very efficiently, and the $\frac{1}{p}\mathbf{S}\mathbf{S}^T$ never has to be explicitly computed, $\mathbf{S}$ can be decomposed into $\mathbf{S} = \mathbf{U}_\mathcal{S}\mathbf{D}_\mathcal{S}\mathbf{V}_\mathcal{S}^T$ which directly results in the sought eigenvalues $\mathbf{D}^2_{\mathcal{S}(ii)} = \lambda_i$ and eigenvectors $(\mathbf{U}_{\mathcal{S}(ij)})_{j=1}^{Np} = \mathbf{u}_i$ of Equation 3.6.

A.5 STATISTICAL MODEL KERNEL

In the reproducing kernel Hilbert space framework, the displacement field u becomes a linear combination of kernels centered at the sampled domain points Equation A.22. If the statistical model kernel is plugged into this framework, for each parameter vector $\boldsymbol{\alpha}$ the displacement field u is a linear combination of sample displacement fields s_i. The weights of this linear combinations are $\boldsymbol{\beta}_{ij}$ Equation A.23.

$$
\begin{aligned}
u(x) &= \sum_i k_S(x, x_i)\boldsymbol{\alpha}_i && \text{(A.22)}\\
&= \sum_i \sum_j (s_j - \mu_S)(x) \otimes (s_j - \mu_S)(x_i)\boldsymbol{\alpha}_i \\
&= \sum_i \sum_j [(s_j - \mu_S)(x)][(s_j - \mu_S)(x_i)]^T \boldsymbol{\alpha}_i \\
&= \sum_i \sum_j (s_j - \mu_S)(x)\boldsymbol{\beta}_{ij}, \\
\boldsymbol{\beta}_{ij} &= [(s_j - \mu_S)(x_i)]^T \boldsymbol{\alpha}_i && \text{(A.23)}
\end{aligned}
$$

A.6 FUNCTIONAL DERIVATIVES

As discussed in Section 2.2 and Section 2.4, the functional

$$\mathcal{J}[u] = \mathcal{D}[u] + \|Au\|^2 \tag{A.24}$$

can be minimized by finding a minimum of the functional derivative with respect to u. Assuming fixed boundary conditions, a minimum bound on $\mathcal{D}$ and A is an invertible linear operator, there exists a minimum $\mathcal{J}[u^*]$ (proofs can be found in Evans [31]). As such, for each variation $\varphi : \Omega \to \mathbb{R}^d$, the function $\epsilon \to \mathcal{J}[u^* + \epsilon\varphi]$ has a minimum at $\epsilon = 0$. It follows that

$$\mathcal{J}'[u, \varphi] := \frac{\partial}{\partial \epsilon}\mathcal{J}[u + \epsilon\varphi]\Big|_{\epsilon=0} = 0 \tag{A.25}$$

if the derivative exists. Applying the fundamental lemma of calculus of variation, we get the strong formulation

$$\mathcal{J}'[u] = 0. \tag{A.26}$$

As example, we provide here the derivation of the functional derivative of the squared loss. Recall, the dissimilarity measure between two images I_0, I_1 with the squared loss

$$\mathcal{D}[u] = \frac{1}{2}\int_\Omega (I_0(x + u(x)) - I_1(x))^2 dx. \tag{A.27}$$

The functional derivative is

$$\mathcal{D}'[u, \varphi] = \frac{d}{d\epsilon}\mathcal{D}[u + \epsilon\varphi]\Big|_{\epsilon=0} \tag{A.28}$$

$$= \frac{d}{d\epsilon}\frac{1}{2}\int_\Omega \frac{1}{C}(I_0(x + u(x) + \epsilon\varphi(x)) - I_1(x))^2 dx\Big|_{\epsilon=0} \tag{A.29}$$

$$= \int_\Omega \frac{1}{C}(I_0(x + u(x)) - I_1(x))\nabla I_0(x + u(x))\varphi(x) dx. \tag{A.30}$$

For simplicity and efficiency, we omit the weighting term C. The strong formulation becomes

$$\mathcal{D}'[u] = \int_\Omega (I_0(x + u(x)) - I_1(x))\nabla I_0(x + u(x)) dx. \tag{A.31}$$

A.7 C2CIRCLE

Matlab code for C and circles

```
m = 512;
[x,y] = meshgrid(linspace(-1,1,m));
R = zeros(m);
R(x.^2+y.^2>0.33^2 & x.^2+y.^2<0.64^2) = 255;
R(x>0 & abs(y)<0.16) = 0;
T = 0.5^2 - x.^2 - y.^2;

figure, imshow(R./255);          % C
figure, imshow(T.*1000./255);    % shaded circle
T(T>0) = 255;
figure, imshow(T./255);          % circle
```

B

TENSOR DECOMPOSITIONS

B.1 ALTERNATING LEAST SQUARES METHOD

The low-rank approximation of H_0 is calculated by minimizing the optimization problem

$$\min_{a_r,b_r,c_r} \|H_0 - \tilde{H}_0\| \quad \text{with} \quad \tilde{H}_0 = \sum_{r=1}^{R} a_r \otimes b_r \otimes c_r. \tag{B.1}$$

The minimizers a_r, b_r and c_r are obtained using the alternating least squares (ALS) method [41]. To this purpose, we introduce a notation to represent a tensor in a matrix form.

Let $H \in \mathbb{R}^{P\times Q\times R}$ be a third-order tensor. By fixing one index the tensor is sliced into two-dimensional sections which have horizontal *(mode-1)*, lateral *(mode-2)* and frontal *(mode-3)* orientation for the indices $\{1, 2, 3\}$ respectively. The mode-n *unfolding* denoted as $H_{(n)}$ concatenates the mode-n slices horizontally to a matrix.

Following Kolda and Bader [50], the CANDECOMP/PARAFAC model can be expressed as

$$H = [\![A, B, C]\!] \equiv \sum_{r=1}^{R} a_r \otimes b_r \otimes c_r,$$

while $H_{(1)} = A(C \odot B)^T$, with $A = (a_1, a_2, \ldots, a_R)$ and likewise B and C. $\odot$ is the Khatri-Rao product (see B.3).

The matrices A, B and C, which minimize Equation B.1 can be calculated by alternately fixing all but one matrix *e.g.* A. This is followed by minimizing

$$\min_A \|H_{(1)} - A(C \odot B)^T\|_F,$$

which has the optimum at

$$A = H_{(1)}[(C \odot B)^T]^\dagger.$$

Using the special property that

$$(A \odot B)^\dagger = (A^T A \star B^T B)^\dagger (A \odot B)^T,$$

where $\star$ is the Hadamard product (see B.3) and $A^\dagger$ the Moore-Penrose pseudo-inverse, the equations can be iteratively solved for A, B and C

$$A = H_{(1)}(C \odot B)(B^T B \star C^T C)^\dagger$$
$$B = H_{(1)}(C \odot A)(A^T A \star C^T C)^\dagger$$
$$C = H_{(1)}(B \odot A)(A^T A \star B^T B)^\dagger.$$

until the values of A, B and C converge. The convergence speed depends on the initialization of the fixed matrices. A common choice for the initialization is to use the Higher-order SVD [30] discussed in Section B.2.

B.2 TUCKER DECOMPOSITION

For the Tucker decomposition, we introduce the following notation to write a tensor

$$\tilde{H}_x = [\![G; A, B, C]\!] \equiv \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} g_{pqr} a_p \otimes b_q \otimes c_r, \tag{B.2}$$

where g_{pqr} are the elements of the core tensor G and P, Q, R are the ranks for each space dimension. In this model the *unfolded* tensor H is represented as

$$H_{(1)} = AG_{(1)}(C \bullet B)^T,$$

where $\bullet$ is the Kronecker product (see B.3). Using the Higher-order SVD algorithm of De Lathauwer and De Moor [30]

$$\min_{G,A,B,C} \|H_{(1)} - AG_{(1)}(C \bullet B)^T\|$$

can be very efficiently minimized by setting A, B and C to the leading left singular vectors of the corresponding mode-n *unfolding* $H_{(n)}$

$$A = U_P^{(1)},\ B = U_Q^{(2)},\ C = U_R^{(3)},$$

where $U_l^{(n)}$ is the matrix consisting out of the leading l singular vectors of $H_{(n)}$ and $G_{(1)} = AH_{(1)}(C \bullet B)^T$.

B.3 MATRIX PRODUCTS

KHATRI-RAO PRODUCT Given a matrix $A \in \mathbb{R}^{m \times q}$ and a matrix $B \in \mathbb{R}^{n \times q}$, the Khatri-Rao product of A and B is the matching column-wise Kronecker product

$$A \odot B = (a_1 \bullet b_1, a_2 \bullet b_2, \cdots a_q \bullet b_q) \in \mathbb{R}^{mn \times q}.$$

HADAMARD PRODUCT Given a matrix $A \in \mathbb{R}^{m \times n}$ and a matrix $B \in \mathbb{R}^{m \times n}$, the Hadamard product of A and B is the point-wise matrix product

$$A \star B = \begin{pmatrix} a_{11}b_{11} & a_{12}b_{12} & \cdots & a_{1n}b_{1n} \\ a_{21}b_{21} & a_{22}b_{22} & \cdots & a_{2n}b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}b_{m1} & a_{m2}b_{m2} & \cdots & a_{mn}b_{mn} \end{pmatrix} \in \mathbb{R}^{m \times n}.$$

KRONECKER PRODUCT Given a matrix $A \in \mathbb{R}^m \times \mathbb{R}^n$ and a matrix $B \in \mathbb{R}^q \times \mathbb{R}^r$, the Kronecker product of A and B is given as

$$A \bullet B = \begin{pmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mn}B \end{pmatrix} \in \mathbb{R}^{mq \times nr}.$$

[7] Thomas Albrecht. *3D statistical shape models of human bones: their construction using a finite element registration algorithm, formulation on Hilbert spaces, and application to medical image analysis*. PhD thesis, University of Basel, 2011.

[8] Thomas Albrecht, Marcel Luthi, and Thomas Vetter. A statistical deformation prior for non-rigid image and shape registration. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[9] Thomas Albrecht, Marcel Lüthi, Thomas Gerig, and Thomas Vetter. Posterior shape models. *Medical Image Analysis*, 2013.

[10] Vincent Arsigny, Olivier Commowick, Xavier Pennec, and Nicholas Ayache. A log-euclidean framework for statistics on diffeomorphisms. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2006*, pages 924–931. Springer, 2006.

[11] Uri M Ascher, Steven J Ruuth, and Brian TR Wetton. Implicit-explicit methods for time-dependent partial differential equations. *SIAM Journal on Numerical Analysis*, 32(3):797–823, 1995.

[12] John Ashburner. A fast diffeomorphic image registration algorithm. *Neuroimage*, 38(1):95–113, 2007.

[13] M Faisal Beg, Michael I Miller, Alain Trouvé, and Laurent Younes. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *International journal of computer vision*, 61 (2):139–157, 2005.

[14] Björn Beuthien, Ali Kamen, and Bernd Fischer. Recursive green's function registration. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2010*, pages 546–553. Springer, 2010.

[15] Christopher M Bishop et al. *Pattern recognition and machine learning*, volume 1. springer New York, 2006.

[16] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194. ACM Press/Addison-Wesley Publishing Co., 1999.

[17] Johan G Bosch, Steven C Mitchell, Boudewijn PF Lelieveldt, Francisca Nijland, Otto Kamp, Milan Sonka, and Johan HC Reiber. Automatic segmentation of echocardiographic sequences by active appearance motion models. *Medical Imaging, IEEE Transactions on*, 21(11):1374–1383, 2002.

[18] Marek Brejl and Milan Sonka. Object localization and border detection criteria design in edge-based image segmentation: automated learning from examples. *Medical Imaging, IEEE Transactions on*, 19(10):973–985, 2000.

[19] Morten Bro-Nielsen and Claus Gramkow. Fast fluid registration of medical images. In *Visualization in Biomedical Computing*, pages 265–276. Springer, 1996.

[20] Nathan D Cahill, J Alison Noble, and David J Hawkes. A demons algorithm for image registration with locally adaptive regularization. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2009*, pages 574–581. Springer, 2009.

[21] Marco Cannone and Grzegorz Karch. On the validity of the picard algorithm for nonlinear parabolic equations. *Proceedings of the Royal Society of Edinburgh: Section A Mathematics*, 135(05):947–958, 2005.

[22] Gary E Christensen and Hans J Johnson. Consistent image registration. *Medical Imaging, IEEE Transactions on*, 20(7):568–582, 2001.

[23] Gary Edward Christensen. *Deformable shape models for anatomy*. PhD thesis, Washington University Saint Louis, Mississippi, 1994.

[24] Tim F Cootes, Mircea C Ionita, Claudia Lindner, and Patrick Sauer. Robust and accurate shape model fitting using random forest regression voting. In *Proc. of 12th European Conference on Computer Vision*, pages 278–291. Springer, 2012.

[25] Timothy F Cootes, Christopher J Taylor, David H Cooper, and Jim Graham. Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59, 1995.

[26] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active appearance models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(6):681–685, 2001.

[27] Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, Clifford Stein, et al. *Introduction to algorithms*, volume 2. MIT press Cambridge, 2001.

[28] Daniel Cremers, Mikael Rousson, and Rachid Deriche. A review of statistical approaches to level set segmentation: integrating color, texture, motion and shape. *International journal of computer vision*, 72(2):195–215, 2007.

[29] Daniel Cremers, Frank R Schmidt, and Frank Barthel. Shape priors in variational image segmentation: Convexity, lipschitz continuity and globally optimal solutions. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–6. IEEE, 2008.

[30] Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. A multilinear singular value decomposition. *SIAM journal on Matrix Analysis and Applications*, 21(4):1253–1278, 2000.

[31] L.C. Evans. *Partial differential equations*. Graduate studies in mathematics. American Mathematical Society, 1998. ISBN 9780821807729.

[32] Bernd Fischer and Jan Modersitzki. Combination of automatic non-rigid and landmark based registration: the best of both worlds. In *Medical Imaging 2003*, pages 1037–1048. International Society for Optics and Photonics, 2003.

[33] Alejandro F Frangi, Daniel Rueckert, Julia A Schnabel, and Wiro J Niessen. Automatic construction of multiple-object three-dimensional statistical shape models: Application to cardiac modeling. *Medical Imaging, IEEE Transactions on*, 21(9):1151–1166, 2002.

[34] James C Gee and Ruzena K Bajcsy. Elastic matching: Continuum mechanical and probabilistic analysis. *Brain warping*, 2, 1998.

[35] Stuart Geman and Donald E McClure. Statistical methods for tomographic image reconstruction, 1987.

[36] Ezequiel Geremia, Bjoern H Menze, Olivier Clatz, Ender Konukoglu, Antonio Criminisi, and Nicholas Ayache. Spatial decision forests for ms lesion segmentation in multi-channel mr images. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2010*, pages 111–118. Springer, 2010.

[37] Ramūnas Girdziušas and Jorma Laaksonen. Gaussian process regression with fluid hyperpriors. In *Neural Information Processing*, pages 567–572. Springer, 2004.

[38] Ulf Grenander and Michael I Miller. Computational anatomy: An emerging discipline. *Quarterly of applied mathematics*, 56(4):617–694, 1998.

[39] Eldad Haber, Stefan Heldmann, and Jan Modersitzki. A scale-space approach to landmark constrained image registration. In *Scale Space and Variational Methods in Computer Vision*, pages 612–623. Springer, 2009.

[40] Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53 (2):217–288, 2011.

[41] Richard A Harshman. Foundations of the parafac procedure: models and conditions for an" explanatory" multimodal factor analysis. 1970.

[42] KA Haskard, SJ Welham, and RM Lark. Spectral tempering to model non-stationary covariance of nitrous oxide emissions from soil using continuous or categorical explanatory variables at a landscape scale. *Geoderma*, 159(3):358–370, 2010.

[43] Matthias Hein and Olivier Bousquet. Kernels, associated structures and generalizations. *Max-Planck-Institut fuer biologische Kybernetik, Technical Report*, 2004.

[44] Darren Homrighausen and Daniel J McDonald. Spectral approximations in machine learning. *arXiv preprint arXiv:1107.4340*, 2011.

[45] Peter J Huber. The 1972 wald lecture robust statistics: A review. *The Annals of Mathematical Statistics*, pages 1041–1067, 1972.

[46] L. Ibanez, W. Schroeder, L. Ng, and J. Cates. *The ITK Software Guide*. Kitware, Inc. ISBN 1-930934-15-7, http://www.itk.org/ItkSoftwareGuide.pdf, second edition, 2005.

[47] Hans J Johnson and Gary E Christensen. Consistent landmark and intensity-based image registration. *Medical Imaging, IEEE Transactions on*, 21(5):450–461, 2002.

[48] Hak Soo Kim, Samuel B Park, Simon S Lo, James I Monroe, and Jason W Sohn. Bidirectional local distance measure for comparing segmentations. *Medical physics*, 39:6779, 2012.

[49] Stefan Klein, Marius Staring, Keelin Murphy, Max A Viergever, Josien PW Pluim, et al. Elastix: a toolbox for intensity-based medical image registration. *IEEE transactions on medical imaging*, 29(1):196–205, 2010.

[50] Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

[51] Ullrich Köthe et al. Vigra-vision with generic algorithms. *Cognitive Systems Group, University of Hamburg, Germany*, 2008.

[52] Rainer Kress, V Maz'ya, and V Kozlov. *Linear integral equations*, volume 82. Springer, 1989.

[53] Dirk-Jan Kroon. Active shape model (asm) and active appearance model (aam). *MATLAB implementation, www: http://www. mathworks. com/matlabcentral/fileexchange/26706-active-shape-model-asm-and-active-appearance-model-aam*, 8:22, 2010.

[54] Rasmus Larsen, Mikkel B Stegmann, Sune Darkner, Søren Forchhammer, Timothy F Cootes, and Bjarne Kjær Ersbøll. Texture enhanced appearance models. *Computer Vision and Image Understanding*, 106(1):20–30, 2007.

[55] Ludovic Lebart. Contiguity analysis and classification. In *Data Analysis*, pages 233–243. Springer, 2000.

[56] Victor Lempitsky, Michael Verhoek, J Alison Noble, and Andrew Blake. Random forest classification for automatic delineation of myocardium in real-time 3d echocardiography. In *Functional Imaging and Modeling of the Heart*, pages 447–456. Springer, 2009.

[57] Michael E Leventon, W Eric L Grimson, and Olivier Faugeras. Statistical shape influence in geodesic active contours. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 1, pages 316–323. IEEE, 2000.

[58] Huiqi Li and Opas Chutatape. Automated feature extraction in color retinal images by a model based approach. *Biomedical Engineering, IEEE Transactions on*, 51(2):246–254, 2004.

[59] Claudia Lindner, S Thiagarajah, JM Wilkinson, GA Wallis, and Timothy F Cootes. Accurate fully automatic femur segmentation in pelvic radiographs using regression voting. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012*, pages 353–360. Springer, 2012.

[60] Zhi-ying Long, Li Yao, and Dan-ling Peng. Fast non-linear elastic registration in 2d medical image. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2004*, pages 647–654. Springer, 2004.

[61] Jyrki Lötjönen, Kari Antila, E Lamminmäki, Juha Koikkalainen, Mikko Lilja, and T Cootes. Artificial enlargement of a training set for statistical shape models: Application to cardiac images. In *Functional Imaging and Modeling of the Heart*, pages 92–101. Springer, 2005.

[62] Huanxiang Lu, Philippe C Cattin, and Mauricio Reyes. A hybrid multimodal non-rigid registration of mr images based on diffeomorphic demons. In *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, pages 5951–5954. IEEE, 2010.

[63] M Lüthi, T Albrecht, T Gass, O Goksel, M Kistler, H Bousleiman, M Reyes, P Buechler, PC Cattin, and T Vetter. Statismo-a framework for pca based statistical models. *The Insight Journal*, 1:1–18, 2012.

[64] Marcel Lüthi. *A machine learning approach to statistical shape models with applications to medical image analysis*. PhD thesis, University of Basel, 2010.

[65] Ján Margeta, Ezequiel Geremia, Antonio Criminisi, and Nicholas Ayache. Layered spatio-temporal forests for left ventricle segmentation from 4d cardiac mri data. In *Statistical Atlases and Compu-*

tational Models of the Heart. Imaging and Modelling Challenges, pages 109–119. Springer, 2012.

[66] R.C. McOwen. *Partial differential equations: methods and applications*. Tsinghua University Press, 1996.

[67] Jan Modersitzki. *Numerical Methods for Image Registration (Numerical Mathematics and Scientific Computation)*. Oxford university press USA, 2004.

[68] David Mumford and Jayant Shah. Optimal approximations by piecewise smooth functions and associated variational problems. *Communications on pure and applied mathematics*, 42(5):577–685, 1989.

[69] James Nolen. Partial differential equations and diffusion processes. Technical report, Technical report, Stanford University. Department of Mathematics, 2009.

[70] Roland Opfer. Multiscale kernels. *Advances in computational mathematics*, 25(4):357–380, 2006.

[71] Xenophon Papademetris, Andrea P Jackowski, Robert T Schultz, Lawrence H Staib, and James S Duncan. Integrated intensity and point-feature nonrigid registration. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2004*, pages 763–770. Springer, 2004.

[72] Nils Papenberg, Janine Olesch, Thomas Lange, Peter M Schlag, and Bernd Fischer. Landmark constrained non-parametric image registration with isotropic tolerances. In *Bildverarbeitung für die Medizin 2009*, pages 122–126. Springer, 2009.

[73] Hyunjin Park, Peyton H Bland, and Charles R Meyer. Construction of an abdominal probabilistic atlas and its application in segmentation. *Medical Imaging, IEEE Transactions on*, 22(4):483–492, 2003.

[74] Xavier Pennec, Pascal Cachier, and Nicholas Ayache. Understanding the "demon's algorithm": 3d non-rigid registration by gradient descent. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI'99*, pages 597–605. Springer, 1999.

[75] A Pintore and CC Holmes. Non-stationary covariance functions via spatially adaptive spectra. *Journal of the american Statistical Association*, 2004.

[76] Allen C Pipkin. *A course on integral equations*. Number 9. Springer, 1991.

[77] Carl Edward Rasmussen. Gaussian processes for machine learning. 2006.

[78] Mikael Rousson and Nikos Paragios. Prior knowledge, level set representations & visual grouping. *International Journal of Computer Vision*, 76(3):231–243, 2008.

[79] Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268, 1992.

[80] Daniel Rueckert, Luke I Sonoda, Carmel Hayes, Derek LG Hill, Martin O Leach, and David J Hawkes. Nonrigid registration using free-form deformations: application to breast mr images. *Medical Imaging, IEEE Transactions on*, 18(8):712–721, 1999.

[81] Marco Ruggeri, Gavriil Tsechpenakis, Shuliang Jiao, Maria Elena Jockovich, Colleen Cebulla, Eleut Hernandez, Timothy G Murray, and Carmen A Puliafito. Retinal tumor imaging and volume quantification in mouse model using spectral-domain optical coherence tomography. *Optics express*, 17(5):4074, 2009.

[82] Matthias Schlachter, Marco Reisert, Corinna Herz, Fabienne Schlurmann, Silke Lassmann, Martin Werner, Hans Burkhardt, and Olaf Ronneberger. Harmonic filters for 3d multichannel data: Rotation invariant detection of mitoses in colorectal cancer. *Medical Imaging, IEEE Transactions on*, 29(8):1485–1495, 2010.

[83] Tanya Schmah, Laurent Risser, and François-Xavier Vialard. Left-invariant metrics for diffeomorphic image registration with spatially-varying regularisation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013*, pages 203–210. Springer, 2013.

[84] Alexander Schmidt-Richberg, Jan Ehrhardt, René Werner, and Heinz Handels. Diffeomorphic diffusion registration of lung ct images. *Medical Image Analysis for the Clinic: A Grand Challenge, MICCAI*, 2010:55–62, 2010.

[85] B. Schölkopf, R. Herbrich, and A. Smola. A generalized representer theorem. In *Computational learning theory*, pages 416–426. Springer, 2001.

[86] Bernhard Schölkopf and Alexander J Smola. *Learning with kernels*. "The" MIT Press, 2002.

[87] Bernhard Schölkopf, Florian Steinke, and Volker Blanz. Object correspondence as a machine learning problem. In *Proceedings of the 22nd international conference on Machine learning*, pages 776–783. ACM, 2005.

[88] Sandro Schönborn, Andreas Forster, Bernhard Egger, and Thomas Vetter. A monte carlo strategy to integrate detection and model-based face analysis. In *Pattern Recognition*, pages 101–110. Springer, 2013.

[89] Carlos OS Sorzano, Philippe Thévenaz, and Michael Unser. Elastic registration of biological images using vector-spline regularization. *Biomedical Engineering, IEEE Transactions on*, 52(4):652–663, 2005.

[90] Aristeidis Sotiras, Yangming Ou, Ben Glocker, Christos Davatzikos, and Nikos Paragios. Simultaneous geometric-iconic registration. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2010*, pages 676–683. Springer, 2010.

[91] Aristeidis Sotiras, Nikos Paragios, et al. Deformable image registration: A survey. 2012.

[92] Florian Steinke and Bernhard Schölkopf. Kernels, regularization and differential equations. *Pattern Recognition*, 41(11):3271–3286, 2008.

[93] J-P Thirion. Image matching as a diffusion process: an analogy with maxwell's demons. *Medical image analysis*, 2(3):243–260, 1998.

[94] Jean-Philippe Thirion. Fast non-rigid matching of 3d medical images. In *Proceedings of the Conference on Medical Robotics and Computer Assisted Surgery (MRCAS'95)*, 1995.

[95] Gavriil Tsechpenakis. Deformable model-based medical image segmentation. In *Multi Modality State-of-the-Art Medical Image Segmentation and Registration Methodologies*, pages 33–67. Springer, 2011.

[96] Ledyard R Tucker. Implications of factor analysis of three-way matrices for measurement of change. *Problems in measuring change*, pages 122–137, 1963.

[97] Matthew A Turk and Alex P Pentland. Face recognition using eigenfaces. In *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91., IEEE Computer Society Conference on*, pages 586–591. IEEE, 1991.

[98] Bram Van Ginneken, Alejandro F Frangi, Joes J Staal, Bart M ter Haar Romeny, and Max A Viergever. Active shape model segmentation with optimal features. *medical Imaging, IEEE Transactions on*, 21(8):924–933, 2002.

[99] J. Vandemeulebroucke, D. Sarrut, and P. Clarysse. The popi-model, a point-validated pixel-based breathing thorax model. In *Proceedings of the 15th International Conference on the Use of Computers in Radiation Therapy–ICCR 2007*, 2007.

[100] Tom Vercauteren, Xavier Pennec, Aymeric Perchant, and Nicholas Ayache. Diffeomorphic demons: Efficient non-parametric image registration. *NeuroImage*, 45(1):S61–S72, 2009.

[101] Grace Wahba. *Spline models for observational data*. Number 59. Siam, 1990.

[102] Yongmei Wang and Lawrence H Staib. Elastic model based non-rigid registration incorporating statistical shape information. In *Medical Image Computing and Computer Assisted Interventation, MICCAI98*, pages 1162–1173. Springer, 1998.

[103] Joachim Weickert, Andrés Bruhn, Thomas Brox, and Nils Papenberg. A survey on variational optic flow methods for small displacements. In *Mathematical models for registration and applications to medical imaging*, pages 103–136. Springer, 2006.

[104] M Yaqub, MK Javaid, C Cooper, and A Noble. Investigation of the role of feature selection and weighted voting in random forests for 3d volumetric segmentation. *IEEE transactions on medical imaging*, 2013.