Outlier Detection for Shape Model Fitting

Inauguraldissertation

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

von

Dana Rahbani

2022

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät

auf Antrag von

Prof. Dr. Thomas Vetter, Universität Basel, Schweiz, Erstbetreuer
Prof. Dr. Volker Roth, Universität Basel, Schweiz, Zweitbetreuer
Dr. Stefan Zachow, Zuse Institut Berlin, Deutschland, externer Experte

Basel, den 14. Dezember 2021

Prof. Dr. Marcel Mayor, Dekan

Abstract

Medical image analysis applications often benefit from having a statistical shape model in the background. Statistical shape models are generative models which can generate shapes from the same family and assign a likelihood to the generated shape. In an Analysis-by-synthesis approach to medical image analysis, the target shape to be segmented, registered or completed must first be reconstructed by the statistical shape model. Shape models accomplish this by either acting as regression models, used to obtain the reconstruction, or as regularizers, used to limit the space of possible reconstructions. However, the accuracy of these models is not guaranteed for targets that lie out of the modeled distribution of the statistical shape model. Targets with pathologies are an example of out-of-distribution data. The target shape to be reconstructed has deformations caused by pathologies that do not exist on the healthy data used to build the model. Added and missing regions may lead to false correspondences, which act as outliers and influence the reconstruction result. Robust fitting is necessary to decrease the influence of outliers on the fitting solution, but often comes at the cost of decreased accuracy in the inlier region. Robust techniques often presuppose knowledge of outlier characteristics to build a robust cost function or knowledge of the correct regressed function to filter the outliers.
This thesis proposes strategies to obtain the outliers and reconstruction simultaneously without previous knowledge about either. The assumptions are that a statistical shape model that represents the healthy variations of the target organ is available, and that some landmarks on the model reference that annotate locations with correspondence to the target exist. The first strategy uses an EM-like algorithm to obtain the sampling posterior. This is a global reconstruction approach that requires classical noise assumptions on the outlier distribution. The second strategy uses Bayesian optimization to infer the closed-form predictive posterior distribution and estimate a label map of the outliers. The underlying regression model is a Gaussian Process

Morphable Model (GPMM). To make the reconstruction obtained through Bayesian optimization robust, a novel acquisition function is proposed. The acquisition function uses the posterior and predictive posterior distributions to avoid choosing outliers as next query points. The algorithms give as outputs a label map and a a posterior distribution that can be used to choose the most likely reconstruction. To obtain the label map, the first strategy uses Bayesian classification to separate inliers and outliers, while the second strategy annotates all query points as inliers and unused model vertices as outliers. The proposed solutions are compared to the literature, evaluated through their sensitivity and breakdown points, and tested on publicly available datasets and in-house clinical examples.
The thesis contributes to shape model fitting to pathological targets by showing that:

- performing accurate inlier reconstruction and outlier detection is possible without case-specific manual thresholds or input label maps, through the use of outlier detection.
- outlier detection makes the algorithms agnostic to pathology type i.e. the algorithms are suitable for both sparse and grouped outliers which appear as holes and bumps, the severity of which influences the results.
- using the GPMM-based sequential Bayesian optimization approach, the closed-form predictive posterior distribution can be obtained despite the presence of outliers, because the Gaussian noise assumption is valid for the query points.
- using sequential Bayesian optimization instead of traditional optimization for shape model fitting brings forth several advantages that had not been previously explored. Fitting can be driven by different reconstruction goals such as speed, location-dependent accuracy, or robustness.
- defining pathologies as outliers opens the door for general pathology segmentation solutions for medical data. Segmentation algorithms do not need to be dependent on imaging modality, target pathology type, or training datasets for pathology labeling.

The thesis highlights the importance of outlier-based definitions of pathologies in medical data that are independent of pathology type and imaging modality. Developing such standards would not only simplify the comparison of different pathology segmentation algorithms on unlabeled

datsets, but also push forward standard algorithms that are able to deal with general pathologies instead of data-driven definitions of pathologies. This comes with theoretical as well as clinical advantages. Practical applications are shown on shape reconstruction and labeling tasks. Publicly-available challenge datasets are used, one for cranium implant reconstruction, one for kidney tumor detection, and one for liver shape reconstruction. Further clinical applications are shown on in-house examples of a femur and mandible with artifacts and missing parts. The results focus on shape modeling but can be extended in future work to include intensity information and inner volume pathologies.

Acknowledgements

I would like to thank Professor Thomas Vetter for guiding this work and supporting me throughout the PhD. Thank you for always making time to discuss progress, share your thoughts and give advice.
This work was possible because of inputs from many. Professor Volker Roth, thank you for your clear feedback after lab meetings that directed my next steps. Dr. Stefan Zachow, thank you for your interest in my work and for quickly providing me with critical feedback and data for the thesis, also for connecting me to Tamaz. Thank you Tamaz Amiranashvili, for your questions that helped me design further experiments. Dr. Beat Schmutz, thank you for your input and discussions about the femur reconstruction experiments. Dr. Lars Ebert, thank you for our collaboration and for inspiring me with SSM application ideas.
GraVis members, thank you for creating an amazing work environment. Andreas Morel-Forster, thank you for the weekly update meetings and the hours of writing and coding clean-up. Marcel Lüthi, thank you for helping me with Scalismo, pointing me to research ideas and encouraging me after lab meetings. Dennis Madsen, thank you for sharing your scripts and teaching me how to improve my presentations and figures. Patrick Kahr, thank you for helping me with Ubuntu and our lively talks. Xolisile Thusini, thank you for staying positive and keeping me hopeful. Jonathan Aellen, thank you for being a good discussion partner. Chunlu Li, thank you for motivating me with your excitement about face image analysis and machine learning conferences. Finally, thank you all for proofreading this document.
To previous GraVis members, thank you for welcoming me to the group. Ghazi Bouabene, Andreas Schneider, Clemens Blumer, thank you for your support with Scala that made my move to a computer science group smoother. Thomas Gerig, thank you for helping me navigate through conferences. Bernhard Egger, thank you for keeping in touch over the years with feedback and discussions. Thomas and Bernhard, thank you for helping

me access the liver data. Adam Kortylewski, thank you for helping me set up the neural network for the CVPR workshop and for pointing me towards part-based models.
I would also like to thank the IT support team as well as Patricia Krattiger for keeping things up and running in the background and making sure my paperwork and open tickets are processed.
Thank you family and friends for your continuous thoughts and prayers and for keeping in touch when I was not. Thank you, mom and dad, for devoting yourselves to my growth even when it meant moving abroad. Thank you Noura for being so understanding in life and research. Thank you Shuk Ha, Wai Hung, Samantha and Joshua for your practical support and encouragement during busy times. I am grateful to have my husband Caleb by my side. Thank you for helping me push through doubt with your patience, faith and love.

Contents

Glossary

SSM Statistical Shape Model

GP Gaussian Process

GPMM Gaussian Process Morphable Model

PBM Part-based Model

GMM Gaussian Mixture Model

BFM Basel Face Model

PPD Predictive Posterior Distribution

EM Expectation Maximization

GAN Generative Adversarial Network

VAE Variational Autoencoder

CNN Convolutional Neural Network

QQ Quantile-Quantile (plot)

ICP Iterative Closest Point

RANSAC Random Sample Consensus

CPD Coherent Point Drift

MSE Mean Squared Error

AD Average Distance

HD Hausdorff Distance

CI Confidence Interval

Chapter 1

Introduction

Shape model fitting is used in medical image analysis tasks such as registration, identification, segmentation, and reconstruction. These applications often rely on a statistical shape model (SSM) for direct shape fitting or for regularization in the background. However, when the target has pathologies, additional steps must be taken to ensure accurate reconstruction results. Pathologies are deviations caused by shape deformations or shape extremes unseen in the example set used to build the model. The problem of pathologies stems from false correspondences that can occur between the reference shape of the model and the given target. Healthy regions do have valid correspondence, while pathological ones do not.
On the one hand, this can be resolved if dense corresponding points are provided by the user and an accurate loss function is used to obtain the reconstruction. The function is built using previous information on the expected pathology - for example, that they generate residuals greater than the expected noise - or a previous segmentation of the pathology that restricts minimization of the loss function to the healthy region. However, both previous knowledge and segmentation labels are pathology- and target-dependent, and therefore require case-specific input from experts. On the other hand, if the reconstruction is available, then the pathology can be detected using a simple alignment. This too requires input from experts for every target in the form of the expected healthy reconstruction.
In comparison, the ideal scenario starts with only the target and SSM and solves for both healthy reconstruction and pathology label map. The ideal scenario would also provide a reconstruction that is not influenced by the outliers, such that the reconstruction in the inlier region is as close as possible

to the one obtained for that target without the outliers [1]. This thesis aims to provide a solution for the ideal scenario. This is accomplished by approaching shape modeling of pathological targets from the viewpoint of outlier detection.

1.1 Noise in GPMM notation

In specific, Gaussian Process (GP) models will be used for inference in this thesis. GP regression has already been proposed and tested for SSM fitting [2, 3]. Gaussian process morphable models (GPMM)s model a shape s as a deformation $u : \Omega \to \mathbb{R}^3$ from a reference shape $\Gamma_R \subset \mathbb{R}^3$

$$s = \{x + \hat{u}(x) | x \in \Gamma_R\}. \tag{1.1}$$

The deformation is modeled as

$$\hat{u} = u + \epsilon \tag{1.2}$$

with deformation $u \sim GP(\mu, k)$ and noise $\epsilon \sim N(0, \sigma^2)$. Common SSMs are represented as a GP by writing μ_{SSM} and k_{SSM} for the mean and covariance function, both estimated from a set of training shapes in correspondence. Deformations sampled from k_{SSM} will be possible shape deformations learned from the training set. Alternatively, kernels can be manually defined to build the GPMM. Manually-defined kernels can also be used to augment the SSM kernel and introduce further flexibility into the model [3].
For any GP, whether defined by μ_{SSM} and k_{SSM} or by analytically defined mean and covariance functions such as a Gaussian kernel, the parameteric low-dimensional formulation used in SSMs can be reached by using the first r basis functions of the KL-expansion of the GP deformation model. A deformation is then the sum of the mean deformation and a linear combination of the eigenvector and eigenvalue pairs (ϕ_i, λ_i)

$$u = \mu + \sum_{i=1}^{r} \alpha_i \sqrt{\lambda_i} \phi_{\mathbf{i}}, \ \alpha_i \sim \mathcal{N}(0, 1). \tag{1.3}$$

Given deformations that map a subset of the reference vertices to those on a target shape, the predictive posterior distribution (PPD) on the remaining vertices is available in closed-form such that

$$\mu_p(x) = \mu(x) + K_X(x)^T (K_{XX} + \sigma^2 I)^{-1} \hat{U} \tag{1.4}$$

$$k_p(x, x') = k(x, x') - K_X(x)^T (K_{XX} + \sigma^2 I)^{-1} K_X(x'). \tag{1.5}$$

The closed form solutions are used to infer the parameters of the marginal distributions in the conditioned GP at each of the remaining vertices. Different points are worth noting in the closed-form PPD.

- The PPD marginal mean $\mu_p(x)$ at a vertex x depends on the mean $\mu(x)$ of the initial GP, the covariance between x and the n observed points $K_X(x) = (k(x, x_i))_{i=1}^n \in \mathbb{R}^{3n\times 3}$, the variance of the training points $K_{XX} = (k(x_i, x_j))_{i,j=1}^n \in \mathbb{R}^{3n\times 3n}$ and the mean free observed deformation values $\hat{U} = ((\hat{u}(x_1) - \mu(x_1))^T, ..., (\hat{u}(x_n) - \mu(x_n))^T)^T \in \mathbb{R}^{3n}$. The mean is in fact a linear combination of the weighted deformations $\hat{U}$, while the covariance is independent of the provided observation values, making the system a linear Gaussian model [4].The mean can be written as a linear combination of kernel functions [5]

  $$\mu_p(x) = \mu(x) + \sum_{i=0}^{n} \alpha_i k(x, x_i) \tag{1.6}$$

  This point is significant for chapter 4 where the PPD is inferred by introducing observations sequentially. The region-growing strategy is only possible because the PPD obtained from sequential observations is the same as the one obtained by conditioning on all the observations together.

- The PPD marginal covariance $k_p(x, x')$ at a vertex x only depends on the location of the obervations -i.e. it is independent of the observed deformation values. To obtain it, only the evaluation of the kernel between the novel point x and the location of the training points on the domain is required [5].

- Deformations that map a subset $\{x_1, x_2, ..., x_n\}$ of the m reference vertices to those on a target shape must be smooth, but there are no further requirements to ensure that they are valid. In cases where correspondence does exist between the reference and target, there exists a set of deformation vectors that are a sample from the learned GP. When applied to the reference mesh, the deformed model sample is a reconstruction of the target. However, this is not guaranteed in cases where the provided correspondence is incorrect due to misalignment or pathologies.

Only noisy observation values are usually available for inference. The PPD formulation above already takes the variance σ^2 of the i.i.d noise ϵ into account. However, outliers are not covered by the noise term. Outliers are those observation values that are not only influenced by noise, but also by pathologies on the target, which means that they are not generated by the shape model nor the Gaussian noise model. The true noisy observation value is not accessible anymore, and using the outlier value instead will generate an inaccurate regression result [6]. It is also not clear which observations are noisy and which are outliers. All this gives rise to the outlier problem. Successful outlier detection makes reconstruction invariant to small deviations imposed on the entire set, or large deviations introduced only locally. Such robust techniques are necessary to address the reconstruction problem [7].

1.2 The outlier problem

Outlier detection has been thoroughly studied in signal processing and for anomaly detection. However, this has not been extended to the shape fitting setting. In this section, we show how outlier detection fits into the shape modeling problem in cases of pathological data, summarized in figure 1.1. Outliers induce a shift in the residuals obtained after reconstruction, making the residuals heavy-tailed. As a result, the noise ϵ no longer follows a Gaussian distribution, which was a necessary assumption for the closed-form PPD in equations 1.4 and 1.5.

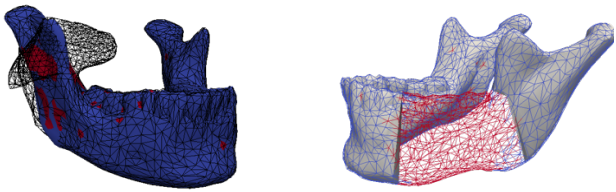


Figure 1.1: Outliers in shape modeling take the form of additional (left) or missing (right) data compared to the model reference.

The robustness of an algorithm to outliers can be studied using sensitivity evaluation and breakdown point analysis. Sensitivity aims to evaluate the influence of one observation point on the regression output. It can be bounded

or unbounded and is often described using an influence function. Breakdown point analysis aims to quantify the number of corrupt observations that can be provided before the algorithm output becomes inaccurate [8].
There are two main schools of thought behind robust GP regression algorithms that handle the outlier problem. The first aims to build more accurate models of the noise. This is necessary because the naive Gaussian noise assumption is not robust due to its fast-decaying tail [9], which means that outliers will lead to errors in the regression output [1, 10]. Since the noise is no longer Gaussian distributed, the closed-form solution cannot be obtained, making approximate inference strategies necessary. Different noise models that take outliers into account have been proposed. For example, GP models have been implemented with noise models that are robust against extreme observations, such as heavy-tailed Student-t distributions or mixture of Gaussians [1, 10]. The Student-t distribution is attractive because its parameters can be adjusted to vary the degree of robustness. Nevertheless, these models require sampling or variational approximation to infer the posterior [10].
In the second approach, the Guassian noise assumption is kept, while the influence of outliers on the regression is decreased. This scenario can make use of 'errors-in-variables' models, in which the observations may be corrupted with additive noise. The influence of corrupted observations is reduced, either through the additional noise or by subsampling the observations. Influence and statistical leverage must together be used to filter out extreme observations when the observation noise is not i.i.d. and corruptions are not sparse. Filtering is therefore determined by both the effect of the observation on the regression as well as its geometric location [11]. An additional classification step or a robust loss function can be used with GP regression to reduce the impact of the outliers on the closed-form solution. The classification step is based on a one-class classifier, which aims to distinguish between the majority "normal" class and minority anomaly class [12].
The two-step alternation between regression and training data classification has been shown to improve regression results in the inlier region. For example, the classification step has been performed using a Student-t distribution for outlier detection, while the regression step that follows uses standard GP regression with Gaussian noise. The classification step consists of a fixed threshold $\alpha = 0.05$ which corresponds to the 5% percentile under the predictive posterior marginal distribution [1]. The proposed approach has a smaller estimate bias in the regressed GP compared to approximate regressions that use the robust Student-t distribution as the noise model.

Other examples that keep the Gaussian noise assumption and filter outliers out come from the statistical shape model fitting literature. In one case, a Bayesian classifier is built to separate occlusion, face, and beard pixels from each other before fitting the Basel Face Model (BFM) to the face region only [13]. Another solution uses a healthy model and an overbite model when fitting a skull SSM to pathological targets. Each model uses its own shape kernel, which keeps the Gaussian noise assumption valid in each model region [14]. This is similar to introducing an outlier model such that observations can be modeled as the sum of the true observation value, the noise, and the bias model [9]. The bias model can be a constant or random vector and must be inferred alongside the GP parameters.
Alternatively, the effect of outliers on the regression can be mitigated by using robust estimators. In the initial GPMM fitting pipeline, a robust Huber loss is proposed to penalize residuals that are larger than a threshold. The threshold can be learned from previous successful registrations or defined by the user [2]. The robust Huber loss is an example of an M-estimator, which weighs the observations according to a weight function [8]. M-estimators have a breakdown point of approximately $1/2$ when the weight function is monotone and bounded. The threshold is called a "kink" point and is needed to determine at which point the residuals are too large to be considered for regression. Another robust estimator is the trimmed estimator, which eliminates the observations with the largest residuals. This sorted estimator is an example of a least trimmed squares LTS-estimator. For example, in GP regression, this can be used iteratively by performing GP regression on all the observations, eliminating the ones with the largest residuals, then performing regression again on the smaller subset [6]. The threshold applied on the residuals can also be estimated from the GP predictive posterior distribution. The predictive posterior mean or variance have been used to determine the most likely inlier observations given previous observations [15]. By combining the two alongside a novelty percentile, the threshold can be set as the sum of the mean and the weighted standard deviation [16].

1.3 Thesis structure

Chapter 1 introduced the problem of outliers in robust GPMM regression and PPD inference and provided the necessary background notation for GPMMs. The rest of the thesis aims to answer the question: how can an accurate predictive posterior distribution be obtained after conditioning a GPMM on observations from a target with outliers? This thesis answers the question

from an outlier detection point of view and is structured as follows:
Chapter 2 explains the different classes of solutions in the literature developed in signal processing to tackle the outlier detection problem. Shape modeling methods specifically developed to deal with pathological targets are then grouped into those classes.
Chapters 3 and 4 cover the two proposed algorithms to obtain the predictive posterior distribution that models the space of possible healthy reconstructions of the pathological target. The first algorithm builds the sampling posterior as obtained from an EM algorithm. The likelihood function is split into two terms to evaluate the foreground and background separately. The EM-like algorithm is inspired from a previous thesis. The proposed method presents an approach to use reconstruction distances instead of pixel intensities in the loss function [13]. The second relies on the closed-form solution of GP regression obtained through sequential Bayesian optimization. A new acquisition function is proposed that takes model certainty as well as the training data into account. The sequential regression ensures that assumptions necessary for closed-form inference are valid in the inlier regions. In each chapter, the relevant background information is first presented, the proposed algorithm is then explained and evaluated followed by a discussion of the underlying assumptions and limitations.
Chapter 5 compares the two approaches, and includes experiments that demonstrate the independence of the fitting strategy from the reference mesh density, the improvement in reconstruction and labeling compared to previous approaches, and the runtime advantage of the sequential GPMM algorithm. The influence of the hyperparameters on the fitting result is studied, based on the landmark uncertainty, the confidence interval used for classification, and the model flexibility. Chapter 6 shows clinical results on publicly available datasets and in-house examples. The generalization ability of the outlier detection algorithms is displayed on different target organs with various pathology types, sizes, and locations. Finally, future work suggestions are listed in chapter 7 before concluding in chapter 8.

Chapter 2

Outliers in Shape Modeling

Organ segmentation approaches rely on labeled data for training discriminative classification algorithms, such as neural networks, or for learning an underlying distribution that characterizes the organ's shape or intensity, such as with Bayesian classification approaches [13, 17]. Both approaches assume that the provided data is sufficient to model the true expected variations of the target. However, when the goal is to segment the organ itself into healthy and pathological regions, the problem significantly changes. With the pathology segmentation problem, the requirements needed for data-based learning or distribution approximation are not always met. Unlike the 'normal' class, the pathology class is not well-sampled and is only accessible through unbalanced training data.
In some cases, the pathology problem can be narrowed down to specific types. This scenario allows for direct modeling of the pathology. Heuristics can then be used to identify and exclude the pathology region. Alternatively, traditional multi-class classification algorithms can be applied through feature extraction or end-to-end learning approaches. For example, naive Bayes has been used for bone tumor diagnosis, but requires examples for 66 tumor diagnosis classes for training [18]. For a novel diagnosis, labeled data and feature definitions for the novel class should be added to the training set, and the training should be repeated. Including novel diagnosis classes is not an unlikely scenario; for example, Dahlin's bone tumors book, which describes different tumor categories, has been updated 6 times since its initial publication date in 1986 to include more examples, evaluation of radiographic variations, and novel cases [19]. This only highlights how broad the pathology problem can be. Therefore, in most cases, constructing an

underlying pathology class from a given dataset is not straight-forward. Some specific pathologies are only rarely observed, others demand high costs and risks to be imaged, while others come in various shapes and sizes. Furthermore, unlike healthy segmentation targets, pathologies cannot be narrowed down to specific image intensities. Pathological data available for training cannot be a representative sample from the full underlying class distribution. Algorithms that solely rely on pathological data for detecting pathologies will have low generalization. This motivates the use of more general characterizations of pathologies. Characterizing the healthy class instead and using it to detect outliers is one promising direction [20, 21].
The problem of detecting outliers in a set of observations is not new. In signal processing, it is known as anomaly or outlier detection. Outlier detection aims to find novel data that are different from previously evaluated datapoints. The core idea behind anomaly detection is one-class classification. From this viewpoint, outlier detection is not based on an outlier model. Instead, it is achieved by learning a threshold on the inlier model [20]. The main challenge in this approach is determining the threshold, since the inliers in practice are not characterized in advance. This thesis views pathology segmentation as an outlier detection problem. For a target surface, e.g. extracted from a medical image volume, a classification of its vertices is desired; healthy points on the surface are inliers, while pathological ones are outliers. This covers both missing data, where outliers are points that exist on a reference shape but not on the target, as well as deformed data and cancerous growth, where outliers are points that exist on the target but not on the reference shape. Pathologies on the target shape can therefore be seen as outliers, which means pathology detection in shape modeling can be addressed using outlier detection algorithms.

2.1 Outlier detection algorithms

There are five categories defined in the review paper [20] in which anomaly detection algorithms fall:

- Reconstruction-based approaches perform regression on the target data, and outliers have high novelty scores in the reconstruction.
- Probabilistic approaches make use of the healthy model distribution, under which outliers have low probabilities and can be identified accordingly.

- Distance-based approaches cluster normal data and define outliers as those far enough from the cluster.
- Information-theoretic approaches compute the information content of the data. Outliers are those points that significantly alter the information content when removed from the dataset.
- Domain-based approaches find a boundary that separates normal data from the outliers, similar to discriminative classification approaches.

Pathology detection in both image and mesh domains can be reformulated in terms of these anomaly detection detection algorithms that perform one-class classification. In this section, pathology detection algorithms from the shape modeling literature are grouped into one of the first three types: reconstruction-based, probabilistic, or distance-based.

2.1.1 Reconstruction-based approaches

Reconstruction-based approaches perform regression on the target data, and outliers have high novelty scores in the reconstruction. In shape modeling, this is equivalent to fitting a healthy data model on a novel target and labeling reconstruction vertices with large distances to the target as pathologies. Since reconstruction residual errors are assumed to be Gaussian i.i.d., a threshold can be set at some standard deviations away from the mean of the residuals, then each residual independently compared to the threshold for classification [22].

Statistical shape models have been used to perform reconstruction of the healthy regions. It has been demonstrated early on that information about the characteristics of healthy regions can be used to obtain a robust shape reconstruction. For example, by using patch decomposition of a mandible bone, it is possible to construct a model of a part of the mandible. The model can be used to perform shape reconstruction while ignoring artifact-prone regions such as teeth [23]. Alternatively, partial active shape models have been proposed to limit shape fitting to specific regions on the target [24]. The regions to be fitted are assumed healthy. Alternatively, the full shape model can be combined with information about the healthy state of the target. Anthropometric landmarks clicked on the pathological target that would have been also found in its healthy state have been used to guide shape model fitting [25], so have characteristics defined on neighboring structures. For example, the known dimensions of an osteosynthesis plate, implanted near a fractured femur, can be used to guide the 3D-reconstruction of the femur

from 2D radiographs [26]. To extract a pathology segmentation, the obtained reconstruction is compared to the pathological target. Detection is therefore based on the reconstruction errors obtained after subtracting the two shapes from each other, under the assumption that a Huber-loss based cost function ensures a robust reconstruction [24]. This approach has been demonstrated also on the image domain, where a healthy cornea shape model has been used to detect anomalies based on pixel residual errors with the target images [22]. Recent approaches utilize generative adversarial networks (GANs) or variational autoencoders (VAEs) trained on healthy datasets [27–29] instead of statistical shape models. In an image reconstruction scenario, given a novel target image with a pathology, the reconstruction should have low pixel/voxel-based intensity residual errors in healthy regions and large errors elsewhere. Different anomaly scores have been proposed in the literature. The anomaly score is a threshold applied on a metric used for quantitative evaluation of segmentation results. Boundary-based approaches that measure contour similarity and region-based approaches that measure area overlap can be used [30]. Some examples of these metrics are the average distance, Hausdorff distance, residual errors, dice score, F1 score and others. For example, the pixel-wise average intensity of healthy data has been used as a threshold, combined with different losses such as the Frechet distance which evaluates the overall similarity between the target and reconstruction using the differences in mean and covariance matrices. Algorithms designed specifically for mesh reconstruction that fall into this category are robust variations of the Iterative Closest Point (ICP) algorithm. Reconstruction errors are trimmed to classify outliers [31] or used to assign weights that filter outlier points out [21, 32]. The threshold in this case is often based on the sorted vertex reconstruction errors.

2.1.2 Probabilistic approaches

Probabilistic approaches make use of the healthy model distribution, under which outliers have low probabilities and can be identified accordingly. The simplest approach fits a statistical shape model to a target. If the reconstruction has a low likelihood under the model distribution, then the target is likely to have pathologies. To ensure the outliers do not influence the full reconstruction, more robust probabilistic approaches have been developed. Previous work in the group argues for using an expectation-maximization like algorithm with sampling to infer the most likely reconstruction distribution. The algorithm alternates between fitting the SSM and classifying model vertices as inliers or outliers using a Bayesian

classifier. The distributions for the different regions are learned during fitting from the pixel intensities in the target or before fitting if the outlier classes are identifiable, such as with beards [13]. In cases where the inlier distribution is not Gaussian, the data can be mapped to another space where confidence intervals can be used as classification thresholds [15]. The mappings can even make it possible to visually examine outliers using QQ-plots [33]. All these approaches aim to learn a threshold that is applied on the class conditional distribution for classification. This threshold is in fact the core behind anomaly detection. Recently, this has been accomplished in CNNs by enforcing the learned class-conditional distribution to be Gaussian, after which the Mahalanobis distance can be used to determine the distance to the class distribution [28, 34]. Nevertheless, the threshold on the distance has to be predetermined according to the task at hand. Furthermore, for every local patch to be evaluated, global features are extracted after removing that patch from the target. This ensures the global features that are extracted are not corrupted in case that patch included outliers, but would be expensive to evaluate in case the target is a surface mesh or if the local patches defined across a 2D image are small and include overlapping pixels [35]. Probabilistic correspondence is another approach which can be used to find most likely inlier pairs and indirectly reject outliers. Rigid and non-rigid registration of 3D surfaces with outliers make use of surface feature descriptors to match regions in correspondence [36, 37] or probabilistic correspondence to choose the most likely matches as in coherent point drift [38].

2.1.3 Distance-based approaches

Distance-based approaches cluster normal data and define outliers as those far enough from the cluster based on Euclidean distance. A similar idea has already been presented in previous work from the group, where two different models are used during shape fitting. In addition to the standard shape model, a GPMM built from a non-stationary kernel models the pathology region [14]. The full model is then built by combining the two using a spatially-varying kernel, which is a weighted linear combination of the two models. The weights are proportional to the Euclidean distance between the model and its corresponding region. Similarly, part-based models (PBMs) split SSMs into parts with a binary occurrence parameter [39]. This solution cannot be directly applied to the pathology problem scenario where the two regions are not known prior to fitting and where the pathologies cannot be defined by a specific deformation model. Another example is the Huber loss

which is used to make robust cost functions for regression. Inliers on the target are assumed to lie within a specific distance of their corresponding point on the reference. The distance is used to define a threshold on the residual errors. Errors below the threshold are important and must be minimized during fitting, while those above the threshold are considered as outliers and ignored during regression.

2.2 Limitations

The categories presented earlier all require a threshold to be defined for outlier classification. Only the underlying meaning behind the threshold differs. Therefore, all of them encounter the same problems when facing general pathologies. Global fitting solutions assume that the shape models used for reconstruction have an ideal balance between model specificity and generalization. This assumption allows the model to perform full target reconstruction, covering both healthy and pathological regions. After reconstruction, a threshold is applied on the residual errors between the target and reconstruction. The classification separates the residuals into two classes: residuals from noise and residuals from pathologies. The first less significant problem is the choice of the threshold. Most are computed at the pixel level using expected intensity residual errors [22, 27, 28]. They are learned from previous healthy reconstruction errors or tuned by the user for different applications. This threshold does not take into consideration the reconstruction quality or make use of model fitting certainty. It also doesn't account for reconstruction errors that can occur due to the pathologies, which leads to the second more important problem of these approaches: using the pathological regions for fitting can corrupt the reconstruction result. This is exactly the problem robust regression tries to solve, whereby it aims to minimize the influence of corrupt data points on the fitting result [1]. Scenarios with low signal-to-noise ratios or with residual errors that vary along the target are problematic for both reconstruction-based and probabilistic approaches. To reduce the influence of outliers, robust estimators must be included, which in turn rely on the threshold which was defined to be used for classification. The threshold is necessary not only for classification, but also for regression to obtain the residuals to be classified, which is a circular problem. Although robust loss functions are available, it is not clear at initialization which to choose because of the broad range of pathologies; they vary not only from one disease to another but also within the same disease.

On the other hand, local to global strategies avoid the problem of corrupt reconstruction of the inlier region. Nevertheless, they face their own set of limitations. The major disadvantage of RANSAC for example is that it assumes that correct correspondences are available, and also relies on random sparse samples instead of faster informed region-growing steps. The thresholds used to control the growth also often face the same problems of the global approach, where a fixed threshold is set by the user or learned from healthy reconstructions without taking into account the influence of outliers on the residual error distributions. Only looking at local patches is also not ideal. It is important to consider the relationship between vertices. In pixel-based anomaly detection, this has been accomplished as local and global feature extraction. An anomaly score is calculated from the local and global anomaly scoring functions [40]. The proposed solution must therefore take outliers into account by minimizing global reconstruction errors, considering neighborhood relationships and adapting the threshold.

Chapter 3

Forward SSM

This chapter presents forward SSM, a shape model fitting pipeline based on the EM-algorithm for targets with pathologies. The motivation comes from previous work [41], in which the Basel Face Model (BFM) was fitted to face images with occlusions by iterating between image segmentation into face and non-face regions and BFM fitting to the face region. The necessity of splitting the target space into regions had already been highlighted [42], in which a background distribution was defined to evaluate pixels in the target image which should not be fitted by the shape model. The pixels to be used in the shape model fitting are described by the foreground distribution model, which is the Gaussian residual error distribution used in traditional shape model fitting. Both foreground and background distributions need to be defined by the user or learned from previous reconstructions with the shape model. The method learns a Bayesian classifier that separates the inliers from the outliers in the E-step and optimizes the model parameters in the M-step. In the medical image analysis context used in this thesis, the foreground distribution is the distribution expected on the inliers, while the background distribution is the one that defines the outlier distribution. Therefore, these distributions are referred to as inlier and outlier region distributions. However, unlike the previous works, this thesis proposes the forward SSM, which enables approximate inference of the predictive posterior distribution (PPD) without requiring assumptions on the outlier region distribution model used for noisy and pathological observations. This chapter explains the contribution using the following outline. In sections 3.1 and 3.1.2, the previous approaches upon which the proposed forward SSM algorithm is based are discussed. The EM-algorithm is first explained, followed by the

limitations of the previous algorithms. Afterwards, in sections 3.2 and 3.3, the proposed algorithm with its necessary assumptions and implementation details are discussed.

3.1 Background

The EM-algorithm is used for estimating the Maximum Likelihood (ML) or Maximum-a-posteriori (MAP) parameters of a distribution in cases of incomplete observations [43]. With incomplete observations, the closed-form solution of the parameters is no longer available, because it is obtained by maximizing the likelihood term which is a function of the full observations. The EM-algorithm resolves this issue by approximating the likelihood term and then maximizing it, and iterating between these two steps until convergence. The first step is called the Expectation (E-)step, while the second is called the Maximization (M-) step. In the E-step, the expected likelihood function is formulated using the available observations and the current estimates of the distribution hyperparameters. In the M-step, the distribution parameters are updated by maximizing the expected likelihood function [10]. Further details can be found in the Pattern Recognition book [44].

3.1.1 EM-like algorithm for SSM fitting

The EM-algorithm is often explained through the Gaussian mixture model (GMM) problem. A GMM is a linear combination of density functions. The goal is to learn the hyperparameters of the density functions. The training data however is not labeled, which means that it is not known which distribution each observation was sampled from [44], which is why the EM-algorithm is necessary. The challenges are deciding the number of density functions and formulating the conditional likelihood function estimate in the E-step. The formulation of GMM inference can be used to understand shape fitting to pathological targets. In pathology detection, the incomplete observations are due to the missing model vertex labels that indicate whether a vertex is an inlier or outlier given a novel target [41]. The number of density functions is therefore only two: inlier and outlier likelihoods. This section explains how to use these likelihoods to obtain the MAP reconstruction of a target.
To implement the EM-like approach in the SSM fitting pipeline, the observations are augmented with an additional set of parameters; labels are

assigned to each vertex on the model reference topology. The binary labels create two classes: an outlier class and a inlier class. The two classes can be seen in figure 1.1. The blue regions belong to the inlier class and the red ones to the outlier class. SSM fitting is restricted to the inlier class to ensure that the outliers do not corrupt the reconstruction. Given a novel target, the vertex locations and labels must be obtained together during registration. The classification and reconstruction steps are together formulated as a maximum a posteriori (MAP) estimation problem. The goal is to find the SSM shape and pose parameters $\boldsymbol{\theta}$ and the point-level label map $\boldsymbol{z}$ that maximize the posterior distribution function given a target surface Γ_T:

$$P(\boldsymbol{\theta}, \boldsymbol{z} \mid \Gamma_T) \propto L(\Gamma_T \mid \boldsymbol{\theta}, \boldsymbol{z}) P(\boldsymbol{\theta}, \boldsymbol{z}) \tag{3.1}$$

The likelihood evaluates the similarity of the SSM reconstruction $\Gamma(\boldsymbol{\theta})$ to the target Γ_T given a specific combination of $\boldsymbol{\theta}$ and $\boldsymbol{z}$, formulated as follows:

$$L(\Gamma_T \mid \boldsymbol{\theta}, \boldsymbol{z}) = \prod_{i \epsilon n} l_{in}(\Gamma(\boldsymbol{\theta})_i, \Gamma_{T\,i})^{z_i} l_{out}(\Gamma(\boldsymbol{\theta})_i, \Gamma_{T\,i})^{1-z_i} \tag{3.2}$$

The n reference vertices are assumed to be independent. Each vertex i is evaluated by either the healthy-region distribution l_{in} or the outlier region distribution l_{out}, determined by the point label z_i. If $z_i = 1$, then l_{in} is used, else $z_i = 0$ and l_{out} is used. The Euclidean distance is used to compare vertex i on the model surface $\Gamma(\boldsymbol{\theta})_i$ and its corresponding point on the target $\Gamma_{T\,i}$. Combining the inlier and outlier distribution models into the likelihood function is necessary to keep the total number of vertices evaluated in the likelihood constant at n. The fixed number of factors evaluated in the likelihood function prevents fitting from converging to a trivial solution such as one obtained by model shrinking [42].

Starting with a surface, the shape parameters $\boldsymbol{\theta}$, label map $\boldsymbol{z}$, and distributions l_{in} and l_{out} are unknown, making optimization intractable. The EM-like algorithm is necessary to infer the posterior distribution. In the E-step, $\boldsymbol{\theta}$ is fixed, the distributions l_{in} and l_{out} are learned, then $\boldsymbol{z}$ is found using Bayesian classification. In the M-step, $\boldsymbol{z}$ is fixed and $\boldsymbol{\theta}$ is inferred as the MAP parameters of the posterior approximated through a sampling strategy. The two steps are explained in more detail in the next paragraphs, and the algorithm is summarized in figure 3.1.

The methodology until this point was developed by previous work [13]. However, unlike previous work, the proposed algorithm aims to find shape deformations in the mesh domain. Therefore, the following changes have to be introduced to the E-step. The goal is to infer the binary label map $\boldsymbol{z}$ defined on the domain of the SSM reference topology. Each of the n vertices

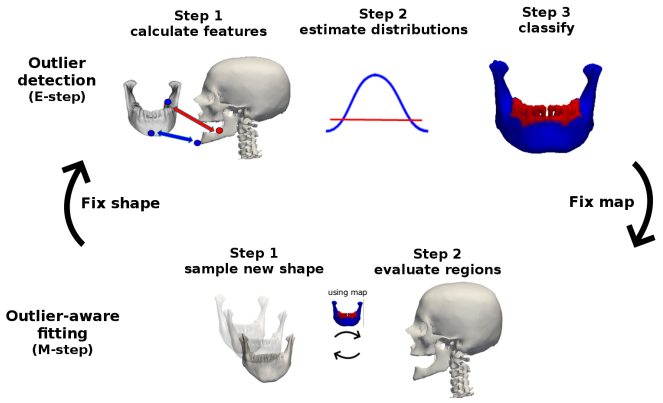


Figure 3.1: Visualization of the EM-based outlier detection and reconstruction algorithm. The input is an unlabeled pathological target surface. The outputs are the reconstruction, the label map and the estimated distributions. They are obtained by iterating between outlier detection (E-step) and outlier-aware fitting (M-step). The label map splits the reference topology into: healthy-region to be used in SSM fitting (blue) and outlier region to be ignored by the SSM (red). This figure and caption were already published in the LABELS workshop at MICCAI 2019 [45].

has a label for one of two classes: healthy-region or outlier region. Starting from equation 3.1, the values of $\boldsymbol{\theta}$ are fixed. The labels which give the MAP solution are then inferred using the following three steps:

- Determine Correspondences: A double-projection method [46] is used. For every vertex of the SSM reference topology, its closest point on the target is found, to which the closest vertex on the SSM is found. The output is a set of bidirectional correspondence pairs. Incorrect pairs are expected because in the beginning the fixed parameters $\boldsymbol{\theta}$ are far from the MAP solution.

- Estimate Distributions: The function l_{in} represents the distribution of residuals which is assumed to be univariate Gaussian. The Gaussian distribution is learned from the current double-projection distances. A uniform distribution for l_{out} is assumed. If some characteristics of the pathology are known and consistent across novel targets, then a more informative distribution can be defined for the outlier region. With the uniform distribution assumption, all pathologies are considered to be equally likely. The likelihood is fixed at the value three standard deviations away from the mean of a healthy distribution, which was learned by fitting to 100 healthy shapes sampled from the SSM.

- Infer Label Map: A point is considered an outlier if its double-projection distance has a higher likelihood of belonging to the outlier region distribution than to the healthy-region distribution. Every z_i is independently inferred by choosing the label corresponding to the larger likelihood value. This is equivalent to maximizing the likelihood function in equation 3.2 with respect to $\boldsymbol{z}$ by classifying bi-directional correspondence distances [47]:

$$L(\Gamma_T, \boldsymbol{\theta} \mid \boldsymbol{z}) = \sum_{i \epsilon n} \sum_{k \epsilon in,out} z_{i,k} l_k(d_i) \tag{3.3}$$

This E-step can be viewed as a probabilistic approach to outlier detection, as described in section 2.1.2 of chapter 2.
To fit the SSM to the target, the SSM parameters $\boldsymbol{\theta}$ that maximize equation 3.1 are found. To do so, the values of $\boldsymbol{z}$ obtained from the E-step are set as fixed. This leaves $\boldsymbol{\theta}$ as the only remaining unknown in the likelihood equation 3.2. The prior on the shape parameters $P(\boldsymbol{\theta})$ is provided by the SSM. With this information, the $\boldsymbol{\theta}$ is found by maximizing the posterior

distribution 3.1 using the likelihood function:

$$L(\Gamma_T, \boldsymbol{z} \mid \boldsymbol{\theta}) = \sum_{i \epsilon n_{out}} l_{out}(d_i(\Gamma(\boldsymbol{\theta}), \Gamma_T) + \sum_{i \epsilon n_{in}} l_{in}(d_i(\Gamma(\boldsymbol{\theta}), \Gamma_T)). \quad (3.4)$$

A sampling approach presented in a previous thesis in this group is used [48]. The posterior distribution is first approximated then the MAP solution is used as the best reconstruction.

3.1.2 Limitations

In the E-step, the outlier region distribution is assumed to be a uniform distribution for generalization purposes [13, 45]. However, when dealing with meshes, two problems arise from this assumption. The first is that all outlier region residuals are equally treated, regardless of their distance to the healthy range. This is not valid in cases where the outliers are due to missing parts on the target. In those cases, a physical outlier point from which the distance to the reconstruction can be measured is missing. The provided method substitutes this value with the distance between the model vertex and a closest point on the target instead of the corresponding point on the target. The second problem comes from restricting the outlier region to follow a distribution. As explained in the oulier detection problem, the distribution of outliers is in most cases intractable. Requiring an outlier model for Bayesian classification therefore necessitates taking assumptions on the type of outliers that are expected, which in turn limits the generalization ability of such a classification scheme.

3.2 Proposed algorithm

Instead of assuming an outlier region distribution, outlier detection can be used instead to learn the outlier threshold during fitting. This section explains the implemented strategy Forward SSM (FSSM), outlined in algorithm 1. In summary, FSSM replaces step 2 in the E-step in figure 3.1 with an outlier detection based classification. The naming FSSM comes from forward-search algorithms, which are explained in the discussion section 3.3.1.
The main assumption taken for this approach is vertex independence.The residual errors at the model vertices are assumed to be independent and identically distributed (i.i.d.). The i.i.d. assumption implies that the likelihood of the residual on each model vertex can be calculated under the inlier region distribution, disregarding the residuals obtained at the

Algorithm 1: FSSM Algorithm

input : GP model $GP(\mu_{SSM}, \Sigma_{SSM})$ as shape prior;
inlier vertices on reference $X^t = \boldsymbol{x_1}, ..., \boldsymbol{x_n}$;
number of samples in M-step N_M;
number of EM steps N_{EM};
target Γ_T;
threshold percentile α;
Output: $\boldsymbol{\theta}$,$\boldsymbol{z}$;
Set labels of X^t in $\boldsymbol{z}$ to inliers;
for N_{EM} **do**
  M-step: choose best $\boldsymbol{\theta}$ after N_M MCMC samples (equation 3.4);
  E-step;
  Get distances of inliers between $d_{X_i^T}(\Gamma(\boldsymbol{\theta}), \Gamma_T)$ for $X_i^T \in X^T$;
  Set threshold t from sorted distances percentile P_α;
  Create empty X^{T+1};
  while $X^{T+1}! = X^T$ **do**
    $X_{new}^T = \Gamma(\boldsymbol{\theta}_j)$,
    $j \in$ neighbors(X^T) and $d_{X_j^T}(\Gamma(\boldsymbol{\theta}), \Gamma_T) <$ t;
    $X^{T+1} = X^T + X_{new}^T$;
  return $\boldsymbol{\theta}$,$\boldsymbol{z}$

neighboring vertices. This is necessary to simplify the joint likelihood functions in the E- and M-steps into the product of independent likelihoods. Nevertheless, the residuals obtained from model fitting are dependent, which is being investigated in another thesis from the group using a correlated noise model.

3.2.1 Outlier detection inspired E-step

The novel threshold in the E-step is calculated from local residual errors of inlier regions used in SSM fitting. The proposed method is therefore a forward-search algorithm [49] in EM-like inference. The label map is initialized from landmarks provided by the user. If the target application should be fully automated, then a landmark detection approach can be introduced for initialization [13, 50]. All points within a certain distance (10mm) of the landmark are assumed to be inliers. To compute the distance on the mesh surface, the method from [51] was implemented in scala and used to generate the geodesic distance map to the landmarks. The influence of the landmark number and uncertainty is evaluated in the experiments sections in chapter 5 and appendix C.
In the E-step, the region-growth evaluates neighbors of inliers. Those which have a residual error smaller than the current threshold are included in the inlier region, after which the next iteration of EM steps is performed. The strategy to include novel points takes mesh connectivity into account instead of ranking all residuals as done in R-estimator techniques. The threshold is learned from the inlier region in the previous fitting step. Instead of relying on a fixed threshold [49], a novel threshold is computed after every M-step and utilized for the next iteration as in dynamic threshold algorithms. A dynamic threshold algorithm does not take a fixed classification threshold as input, but instead updates it along with model fitting and provides it as an output [52]. The dynamic threshold is often computed as an average from previous observations, which, based on the central limit theory, will be Gaussian distributed even if the observations themselves are not sampled from a Gaussian distribution [52]. The Hausdorff distances between target and reconstruction are used to learn the current threshold. Section 3.3.2 discusses this choice and compares it to using the average distance and likelihoods instead.

3.2.2 M-step

The M-step approximates the PPD of obtaining the target given the SSM parameters and the outlier map. The distribution can later be used to estimate the certainty of the outlier detection or to get the best reconstruction, which is the MAP solution of the obtained distribution. SSM fitting is performed using the provided map as input $\boldsymbol{z}$ in equation 3.2: all significant points from the landmark initialization are in l_{in}, while all others are in l_{out}. The outlier likelihood is kept as a uniform distribution.

3.3 Discussion

3.3.1 Comparison to Forward Search Algorithms

In computer graphics, a forward-search algorithm starts from a subset of the data that excludes outliers then includes more points into this subset based on a chosen metric. The idea is to include only valid points to do a local surface reconstruction instead of performing a reconstruction based on all observations then using it to exclude outliers. Forward-search is similar to the seeded region growing algorithm first presented to perform semantic image segmentation [53], but aims to perform piece-wise smooth surface reconstruction from point clouds instead of segmentation [49].
To detect the onset of regions with outliers, jumps in residual errors can be used. This information drives a forward-search algorithm to include points with residual error values similar to that of their neighbors, improving the reconstruction estimate in a local-to-global approach [49]. Schall et al. [54] perform surface reconstruction from point cloud data by estimating the density function of the surface from which the data is assumed to be sampled. A point can then be assigned a probability of belonging to the estimated surface based on its distance to a locally fitted plane. It can also be moved around to increase the likelihood of observing the full point-cloud, which is often obtained through a noisy scanning process. This is a straight-forward approach that can handle sparse outliers. However, to deal with large-amplitude clustered noise as would be the case of pathologies, the size of the local neighborhood used to obtain the plane must be adjusted. Similarly, Wang et al. [55] propose a strategy based on residual fitting errors to detect non-isolated outliers that cause smooth deformations on the target surface. The local neighborhood of every point is estimated by a quadratic surface, then a voting strategy is used to determine if the point is an outlier or not based on its neighbors' distances to the local fit. Here, the size of the

local neighborhood and the density of points within it can significantly influence the detection results: smaller regions and higher densities will underestimate the number of outliers. Another strategy views the problem as a segmentation or clustering one. Local primitive fitting can be used to split a mesh into parts which fall into the same primitive category [56], and similarly on point cloud data local surface variation features can be used for k-means clustering into outlier or inlier regions.
The common assumption underlying these approaches is that local surfaces of the target can be approximated by simple primitives or fitted with robust polynomial regression techniques. However, some applications call for more complex shapes to be scanned. In those cases, robust local fitting approaches would require small local scales in order to successfully fit them to well-defined primitives or planes, because even having a couple of outliers in a local region can cause the least-squares plane fitting to fail [49]. Small local scales would again bring up the problem explained above of underestimating outliers [55]. Another assumption that these approaches build on is that the surface can be reconstructed with hundred percent certainty, which is often not the case with SSM reconstructions. Surface reconstruction is an ill-posed problem where many plausible reconstruction solutions can be provided for just one point cloud set. This makes it necessary to associate each reconstruction with a certainty value. Maps that highlight the spatial probability of locating a point on a surface or the probability of obtaining a full surface consisting of the observed pointset can be introduced [57], but these too assume that robust local linear reconstructions exist and do not make use of priors on geometry.
The proposed solution is an example of a forward-search algorithm which makes use of a geometry prior. Local fitting results are used to guide the region-growing and therefore influence global shape reconstruction with the SSM. The SSM in the background improves robustness by limiting the allowed deformations.

3.3.2 Residual Error Monitoring

Different metrics can be used to detect jumps in residual errors when performing local-to-global SSM fitting. This section evaluates the Hausdorff distance, average distance, and likelihood of the reconstruction obtained from the SSM shape parameters as metrics. It also discusses whether a forward-search is necessary by comparing fitting to local regions obtained by: uniformly sampling a subset of points across the surface, outlier-agnostic region growing by always adding the neighboring vertices within geodesic

distance d to the inlier subset. The number of points used for fitting is matched in both cases.
Label maps generated from each approach are depicted on the reference mesh in the first two rows of figure 3.2. Blue indicates the local region to be fitted, and red indicates the region excluded from fitting. The number of vertices used in fitting increases from left to right, which is visible by the increase in the size of the blue region. The plots in the third row of the figure summarize the change in the value of the metric used to compare the target to the reconstruction as a function of the inlier region. The x-axis quantifies the size of the blue region. It is described in terms of the geodesic distance to the initial landmark on the right condyle, seen in the first row of the figure. As the set of used vertices increases, the distance metrics are expected to decrease if the fitted region is healthy, accompanied by an increase in the log-likelihood value. A jump in the metric values towards the opposite direction would indicate a pathological region on the target. The ground truth pathology region is located between the vertical dotted lines in the plots. The jump can be detected in the correct range when using the Hausdorff distance metric. This is more significant when using the growth strategy based on region-growing (green) instead of uniform sampling (red). Therefore, a percentile of the Hausdorff distance and the region growing approach are used to calculate the threshold and drive the local-to-global fitting strategy in the forward SSM.

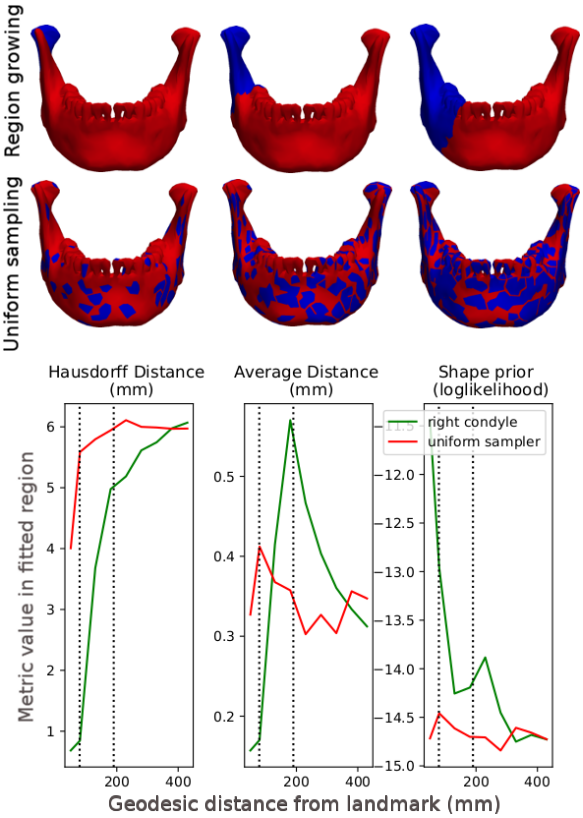


Figure 3.2: Residual error monitoring as the evaluation region in blue increases. The plot shows the metric value versus growth strategy: region growing (green) and uniform sampling (red). The outlier region is located between the dotted lines. The jump in Hausdorff distance with the region growing strategy accurately localizes the outlier region.

Chapter 4

Sequential GPMM

This chapter presents the sequential GPMM algorithm. Sequential GPMM is a shape modeling pipeline based on sequential Bayesian optimization for targets with pathologies. The proposed sequential GPMM algorithm enables closed-form inference of the predictive posterior distribution (PPD) even when the target structure has pathologies.
In sequential Bayesian optimization, inference occurs by introducing observations one after the other [58]. Optimization occurs as an iterative process, cycling between choosing an observation to be used in inference and then updating the inference with that observation. In every iteration, the number of observations used in inference increases. The order in which the observations are introduced is determined through an acquisition function, explained in detail in section 4.1.1. The acquisition function is optimized to choose the observations that should be used in the next inference step. For example, one criterion that can be used to define an acquisition function is the remaining uncertainty in the model after inference. The acquisition function is therefore used to find the observations that, if chosen for inference, would minimize the remaining uncertainty in the PPD. For sequential GPMM, the fitting criterion is minimizing the number of outliers used in regression. Therefore, the main contribution of the proposed sequential GPMM algorithm is the robust acquisition function, which will be explained in detail in section 4.2.3.
In most shape modeling applications, a set of initial landmarks are provided to perform initial rigid alignment of the model and target. Landmarks, either the ones provided for alignment or others, can also be used to obtain an initial posterior distribution from the GPMM. The sequential GPMM algorithm

uses the initialization landmarks as seeds for a region-growing algorithm. The growth occurs successively by including vertices for which the model is confident that they have a corresponding point on the target surface, while avoiding the vertices which for which the model has low confidence. The latter are labeled as pathologies. For this, a classification function must be learned and integrated into the fitting pipeline as the acquisition function. While previous approaches only rely on the model certainty from the predictive posterior variance to calculate the threshold, the proposed acquisition function also takes the inlier reconstruction quality and the values of potential query points into account. One assumption is that the shape model prior is built from healthy data and is able to generalize to fit to novel instances. Another assumption is that the shape model posterior, obtained by conditioning the prior on inlier observations, can make reliable predictions about the mean and covariance remaining in the model. Unlike FSSM, the sequential GPMM obtains the mean and covariance using the closed-form PPD solution instead of approximating it with sampling.

4.1 Background

Sequential Bayesian optimization methods are used to infer model parameters if the cost of obtaining observations is high or if the size of the training data is too large to allow for direct training of the model on all the data [4]. Also known as active learning, sequential learning, or online data selection, the main idea is to introduce observations one after the other when inferring the model parameters.
The structure of a Bayesian Optimization algorithm is outlined in table 4.1. The algorithm alternates between obtaining an observation and performing inference using the collected observations. Introduced observations are called query points. They are introduced in an order determined by an acquisition function, which quantifies the model improvement obtained from introducing a specific observation. For one-class classification, the different steps are termed as following. The 'learning scenario' defines the model and initial conditions. It is summarized in the input row in table 4.1 below. The 'base learner' acts as a binary classification function, and the 'query strategy' determines the points to be evaluated by the base learner [12]. The base learner and query strategy together formulate the acquisition function [59], shown as in step 2 of table 4.1 below. The acquisition function will be explained in more detail with some examples from literature in section 4.1.1. In every iteration, the observation which maximizes the acquisition function

Table 4.1: Outline of sequential Bayesian optimization algorithm and its notation for shape fitting with GPMMs.

	Sequential Bayesian Optimization	Sequential GPMM	Notation
Input	Inlier observations and regression model	Landmarks and GPMM	X_n^0, GP(μ_{SSM}, Σ_{SSM})
Step 1	Regression on inliers	GPMM conditioning on landmarks	GP(μ_{PPD}, Σ_{PPD})
Step 2	Choose next query point	Choose healthy inlier neighbors	X_{n+q}^{t+1}
Output	Regressed function parameters, inliers	Shape reconstruction, inlier landmarks	θ^t,X^t

is chosen as query point and introduced to update the model parameters. The acquisition function determines the order with which new query points are introduced. However, this order does not influence the final regression output. Furthermore, the final regression output is also not influenced by whether query points are introduced simultaneously or sequentially. Sequential updates of GPMMs are possible because GPs can be formulated as a linear combination of kernels, and the PPD mean is itself a linear combination of the deformations, as shown in equations 1.4 and 1.6. The PPD mean and covariance depend on the query point locations, but not their values. Therefore, sequential updates do not bias the intermediate results or final output [5, 58]. Therefore, although batch and sequential optimization are coded differently, they result in the same posterior distribution at the end. Obtaining the model parameters $\boldsymbol{\theta}$ with sequential Bayesian updating can be formulated with Bayes' rule as follows:

$$P(\boldsymbol{\theta}|\boldsymbol{X_n}, X_q) \propto P(X_q, \boldsymbol{X_n}|\boldsymbol{\theta})P(\boldsymbol{\theta}), \tag{4.1}$$

where q and n indicate new and previous query points used for model conditioning. Since the query points are conditionally independent given the

model parameters, then

$$P(\boldsymbol{\theta}|\boldsymbol{X_n}, X_q) \propto P(X_q|\boldsymbol{\theta})P(\boldsymbol{X_n}|\boldsymbol{\theta})P(\boldsymbol{\theta}) \tag{4.2}$$

This implies that

$$P(\boldsymbol{\theta}|\boldsymbol{X_n}, X_q) \propto P(X_q|\boldsymbol{\theta})P(\boldsymbol{\theta}|\boldsymbol{X_n}) \tag{4.3}$$

Therefore, the posterior obtained when training on $\boldsymbol{X_n}$ can be used as the prior when training on $\boldsymbol{X_q}$. For a detailed explanation, please refer to chapter 4 of the book [60].
Active learning with sequential Bayesian regression and one-class classification with Gaussian processes can be seen as two sides of the same coin. The former aims to find query points that can provide the most accurate inference output at minimum cost. The cost is due to the difficulty of obtaining observations or the large size of the dataset used in training. The latter aims to find query points that can provide the most accurate inference output given a two-class training dataset. The query points must all come from the correct class, which necessitates classification of the dataset before or during inference. The sequential GPMM algorithm was developed for this classification task, and as such any assumptions that it is built on are made without violating this requirement. The main difference comes from the purpose of the query point. In the former, the query point attempts to provide the most information for inference. Minimizing the number of query points is therefore of interest. In the latter, the number of query points used is not important, as long as they are all inliers of the target class, which again is the contribution of the sequential GPMM. Nevertheless, both approaches rely on acquisition functions to choose the correct query points, which will be discussed next.

4.1.1 Acquisition Functions

The acquisition function quantifies the effect of introducing a specific observation x on the inference [1]. In every iteration, the observation which optimizes the acquisition function is introduced. Different acquisition functions have already been developed:

- Acquisition functions based on information gain: The query point is selected as the one that maximizes information gain about the model parameters. The change in entropy of the distribution on the model parameters is used to measure the information gain, where a bigger drop represents a larger information gain since the distribution variance

drops. The cross entropy between the two distributions can also be used, which instead quantifies the improvement in the mean of the distribution. It has been shown that both correspond to setting the query point to be the observation with the highest inferred remaining variance in the posterior distribution, known as Mackay's active learning criterion (ALM) [61]. To make local query point choices, local regions can be defined in which the fitting quality must be more accurate than in others [58]. In GP regression, the posterior and predictive posterior variance are available from the closed-form solution or by approximation methods. The observation with the highest posterior variance is the one that will be chosen using ALM. The one that would cause the most significant drop in overall posterior variance will be chosen as query point using Cohn's criterion (ALC) [61]. Variance has also been used in one-class classification with GPs [15]. The combination of variance, mean and information gain has also been proposed [62].

- Acquisition functions generally defined through expected improvement functions [1]: The expected improvement function is formulated as a weighted sum of the PPD mean and covariance at a query point. It has a closed-form solution for GP regression [63] [64]. The weights determine the proximity of query points to previous observations, such that reducing the weight of the mean to zero results in the variance-based solution of the information gain acquisition functions.

4.1.2 Limitations

Compared to traditional sequential Bayesian optimization, shape modeling does not face the problem of expensive observations. Observations are obtained by a correspondence function, which can be defined as a simple closest point in space according to Euclidean distance or as a more detailed surface feature match. The cost is the same for all observations and batch optimization is possible.
However, GPMM regression to pathological targets faces corrupt observations which should not be used in inference. The acquisition functions above assume that all observations and chosen query points are not corrupted [61] and use the predictive posterior variance. While the variance relies on the underlying covariance function used to build the GPMM, it does not consider the observation values, as explained using equation 1.5. This implies that the model confidence about the marginal mean at a point will be large if that point has high correlation to previous observations or if many of

its neighbors have already been observed, regardless of the value of the observation at that point. Therefore, model confidence cannot be used on its own as a measurement of generalization error when there are corrupt observations [61]. Shape modeling in the presence of outliers is one such scenario; false observation values are not equal to the ground truth deformation with Gaussian noise modeled in equation 1.2. A rejection function has to be separately defined to be used alongside the variance for error evaluation [61]. An acquisition function that considers errors in the observations is necessary for fitting GPMMs to pathological targets. Furthermore, the information gain approaches assume that the probability distributions described by the model are correct [58]. That is, for a GP, the remaining variance in the posterior predictive distribution is accurate. The posterior variance however is only as correct as the model that is used to obtain it [61], and in case of the GPMM only a low-rank approximation is used. This thesis resolves this issue through an acquisition function that accounts for inaccurate model variance by learning a threshold during model fitting. Alternatively, model accuracy can be improved, which is currently being investigated in the group through a correlated noise model and will be discussed further in the future work chapter 7.

4.2 Proposed Algorithm

Sequential GPMM regression extends standard sequential Bayesian optimization algorithms to be used in shape modeling with pathological targets. The algorithm consists of a surrogate function and an acquisition function. The GPMM is given and used as the surrogate function. The main contribution of this thesis is to propose and add a robust acquisition function. The relationship between sequential Bayesian optimization steps and the sequential GPMM algorithm is summarized in table 4.1, along with the relevant notation. Figure 4.1 visualizes the sequential GPMM regression and classification steps using a 1D function. The function $f(x)$ (dotted black) represents a target shape, from which noisy data is available. As the number of observations increases, the PPD mean (solid green) is updated and its variance decreases. A threshold is computed from the Mahalanobis distances of previous observations (blue) and used to perform outlier detection. Unlike the confidence intervals (green shades), the threshold takes into account not only the posterior variance but also the observation values by using their Mahalanobis distance to the current mean. It is used to determine the query points (yellow) which will be added as observations in the next iteration of

GP regression. The outliers (red) are not labeled in advance, but only shown here in red to distinguish them from the correct observations (black). Some of the outliers will be incorrectly classified as query points. These are marked with stars (red). However, in the mesh scenario, query points are not searched for on the entire domain. Instead, only neighbors of observations are evaluated at every iteration. This makes it possible to avoid the outlier query points.

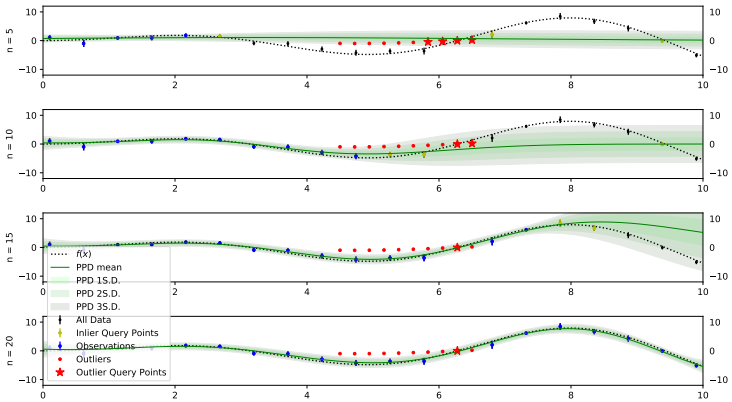


Figure 4.1: Visualization of the sequential GPMM algorithm using GP regression over a 1D continuous domain.

Algorithm 2 summarizes the standard Bayesian optimization pipeline. Starting from a surrogate function and n input observations, the goal is to find the function parameters θ. The observations $\boldsymbol{x}$ are written in bold to indicate that they can be vectors, while the observation values y in this example are scalars. Optimization is accomplished by regressing the surrogate function on N observations, a subset of the total available observations which fulfills the criteria described by the acquisition function $J(\boldsymbol{x})$. In a standard sequential Bayesian optimization algorithm, it is often desired to have a small number of query points for speed or cost purposes, so N is always less than the total number of available observations. The iterative algorithm starts by finding $\boldsymbol{\theta_n}$, the function parameters obtained from the initial n observations. The function is then used in the acquisition function to determine the next query point $\boldsymbol{x_{n+1}}$. The novel query point is included with the initial n observations, after which the process repeats itself until the total number of observations reaches the desired value N or another stop condition is

reached. The stop conditions are defined for the application at hand as a threshold on the objective function. For example, if the information gain is used, then the stop condition will be reached when the information gain of including a novel observation drops below a desired threshold.
The described algorithm is extended to be used in shape model fitting. The result is the proposed sequential GPMM algorithm, outlined in algorithm 3. The observations are the model vertices and their values are the deformation vectors at these vertices. The deformation vector at each vertex has a norm equal to the distance between the vertex and its corresponding point on the target. The deformation vector displaces the vertex to the position of its corresponding point. A novel instance from the shape model is obtained by applying the deformation vectors to all the model vertices. This is applied on all the reference vertices The additional variables needed for shape modeling are marked with a $\boxed{\star}$ symbol. The additional input is a correspondence function, which is necessary to determine how the observation values are obtained. The additional output is a binary label map, which is built by labeling the reference mesh vertices that were used as query points X^t as inliers. The boxes indicate the three major blocks that are introduced in sequential GPMMs to make the sequential Bayesian optimization pipeline robust. The introduced blocks are discussed in detail in the next subsections.

Algorithm 2: Standard algorithm for Bayesian optimization/active learning/sequential learning

input : surrogate function $f_{\theta_\mathbf{0}}(\boldsymbol{x})$;
inputs $X_n = \boldsymbol{x_1}, ..., \boldsymbol{x_n}$;
design size N;
acquisition function $J(\boldsymbol{x})$;
Output: regressed surrogate function parameters $\boldsymbol{\theta_N}$;
repeat
1. get observations $D_n = ((\boldsymbol{x_1}, y_1), ..., (\boldsymbol{x_n}, y_n))$;
2. fit surrogate to observations $\boldsymbol{\theta_n} = argmax_{\boldsymbol{\theta}}(f_{\boldsymbol{\theta}}(\boldsymbol{x})|D_n))$;
3. choose query point that maximizes acquisition function $\boldsymbol{x_{n+1}} = argmax_{\boldsymbol{x}}(J(\boldsymbol{x}, f_{\theta_n}(\boldsymbol{x}))$;
4. add query point to observations $X_{n+1} = \boldsymbol{x_1}, ..., \boldsymbol{x_n}, \boldsymbol{x_{n+1}}$, making $n == n + 1$;

until $n == N$;

Algorithm 3: Skeleton algorithm for Sequential GPMM regression

input : surrogate function: GP model $GP(\mu_{SSM}, \Sigma_{SSM})$;
inputs: inlier vertices on reference $X^t = \boldsymbol{x_1}, ..., \boldsymbol{x_n}$ [1];
design size: total number of reference vertices N;
acquisition function: classification criterion $J(\boldsymbol{x})$;
target: Γ_T with correspondence distance function d_c [⋆];
Output: Posterior GP with regressed parameters $\boldsymbol{\theta^t} = (\boldsymbol{\mu^t}, \boldsymbol{\Sigma^t})$;
observations X^t [⋆];
repeat
1. get observations $D^t = ((\boldsymbol{x_1}, \boldsymbol{y_1}), ..., (\boldsymbol{x_n}, \boldsymbol{y_n}))$ using $\boldsymbol{y} = \arg\min_y d_c(\boldsymbol{x}, \boldsymbol{y}, \Gamma_T)$ [2];
2. fit surrogate to observations $\boldsymbol{\theta_n} = (\mu_{PPD}, \Sigma_{PPD})$ using PPD equations 1.4 and 1.5.;
3. choose query points that satisfy criterion $\boldsymbol{X^{t+1}} = \{\boldsymbol{x}|J(\boldsymbol{x})|\boldsymbol{\theta^t} \neq 0\}$ [3];
4. add query points to observations $X^{t+1} = \boldsymbol{x_1}, ..., \boldsymbol{x_n}, ..., \boldsymbol{x_{n+q}}$, where q is the number of obtained query points;

until $\mathbf{X}^t == \mathbf{X}^{t+1}$;

4.2.1 Landmark Initialization

inlier vertices on reference as inputs $X^t = \boldsymbol{x_1}, ..., \boldsymbol{x_n}$ [1]

The landmarks on the reference mesh shape are set as inliers on the initial label map. The provided landmarks are assumed to exist as inliers on the target shape. To account for errors in landmark placement, regression is performed with the assumption of Gaussian noise on the input, the influence of which is analyzed in the experiments section in figure 5.8. The target is initially rigidly aligned to the model based on a set of initial inlier pairs where y_i is the landmark on the target and x_i is its corresponding landmark on the reference shape. An initial binary label map $\boldsymbol{z}$ is built, with length equal to the number of model reference vertices and elements that indicate the classification result of each vertex. Therefore, $z_i^{t=0} = 1$ if x_i is one of the landmarks else $z_i^{t=0} = 0$. The number of landmarks and their locations in this thesis are clicked by the user on the reference shape and are the same ones used for the initial rigid alignment of the model. The influence of the number of landmarks and accuracy on the results is evaluated in section 5.2.2. Automatic methods that can be used to replace the manually clicked landmarks are discussed in the future works chapter 7.

4.2.2 Correspondence Function

get observations $D^t = ((\boldsymbol{x_1}, \boldsymbol{y_1}), ..., (\boldsymbol{x_n}, \boldsymbol{y_n}))$ using $\boldsymbol{y} = argmin_y d_c(\boldsymbol{x}, \boldsymbol{y}, \Gamma_T)$	2

The observation values are obtained based on a predefined correspondence function. The correspondence function is used to obtain y_i^t, by matching a vertex on the mean of the current GPMM posterior $\mu_p^t(x_i)$ to the target Γ_T. The mean of the GPMM posterior is used because it is the MAP solution of the reconstruction of the target. This implies that the estimated correspondences will change whenever the GPMM posterior is conditioned on more observations. The correspondence function used in this thesis is based on the dot product of the normal vectors:

$$d_c(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{n_x}, \boldsymbol{n_y}, \Gamma_T) = \boldsymbol{n_y}.\boldsymbol{n_x} \tag{4.4}$$

From the points on the target that lie within a specific distance r from the reference vertex, the observation is the one that shares the most similar normal vector direction to that of the reference vertex. The distance r is heuristically set to $10mm$. Because the correspondences are estimated with every posterior update, there are less errors in correspondence as the reconstruction gets closer to the target shape, similar to the correspondence estimation approach in iterative closest points [65]. The observation value is therefore

$$\boldsymbol{y_i^t} = argmin_{\boldsymbol{y}} d_c(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{n_x}, \boldsymbol{n_y}, \Gamma_T) \forall \boldsymbol{y} \in \Gamma_T \text{with} ||\boldsymbol{x} - \boldsymbol{y}|| < r \tag{4.5}$$

The normal vector condition is introduced to prevent incorrect matches of points on the outer mesh surface with those on the inner surface. This is the case for targets with two sides such as the cranium or the femoral shaft. Furthermore, a double projection approach is used [47]; first, the corresponding points of each vertex on the model reference is found on the target, and second, the corresponding points of the obtained points on the target are found on the model. The superscript t indicates that the variable is updated with every novel PPD obtained from inserting novel query points to the sequential GPMM regression. This block can be replaced by a case-specific correspondence function or a descriptor-based approach for challenging shapes. This was not evaluated in this thesis but can be covered with alternatives similar to the automatic landmark generation approaches discussed in the future works chapter 7.

4.2.3 Acquisition function

The acquisition function is used to determine the next query points. It is built from (1) the Mahalanobis distance of the query points to their marginal distribution from the PPD (2) the Mahalanobis distance of previous observations to the marginal conditional distribution. The Mahalanobis distance is used because, unlike the Euclidean distance, it takes the model uncertainty into account. The dynamic threshold therefore makes use of both the distance to the mean and the uncertainty under the posterior. The acquisition function allows filtering of false observation values that are actually outliers without requiring input outlier labels or a predetermined classification threshold. A nearest neighbor query strategy is used, where only neighbors of previous observations are evaluated.

choose query points that satisfy criterion $\boldsymbol{X^{t+1}} = \{\boldsymbol{x}\|J(\boldsymbol{x})\|\boldsymbol{\theta^t} \neq 0\}$	3

The PPD has multiple known characteristics [4] that are used to build the acquisition function:

- First, given the smooth Gaussian kernel and low-rank approximation as explained in section 1.1, the inferred variance is smaller for points on the domain which are closer to the observed points $\{\boldsymbol{x_1}, ..., \boldsymbol{x_n}\}$. This implies that, under the PPD, the model is more certain about the mean at $\boldsymbol{x_{n+1}}$ if it lies close to any of the observation locations $\{\boldsymbol{x_1}, ..., \boldsymbol{x_n}\}$ and less certain about faraway regions.

- Second, under the same conditions, the variance remaining at any point under the PPD decreases as the number of observations increases. To counter the variance drop [15], the acquisition function is updated as more observations are included in the PPD. This can be seen as increasing the threshold on model confidence below which query points are updated to inliers. The following example illustrates this point. In the case of estimating the Maximum-Likelihood solution of a Gaussian distribution, the mean value is updated with the introduction of a novel observation x_{N+1} as derived in section 2.3.5 of [4]:

$$\mu_{N+1} = \mu_N + \frac{1}{N+1}(x_{N+1} - \mu_N) \tag{4.6}$$

  The change in the value of the mean decreases as the number of observations increases. As more observations are provided for training, the model updates the posterior distribution less [4, 60]. The rate of change of the acquisition function threshold therefore cannot be fixed, but must account for the total number of observations.

- Finally, the order of introducing the observations does not influence the final predictive posterior distribution; conditioning on $\boldsymbol{x_1}$ then on $\boldsymbol{x_2}$ gives the same predictive posterior distribution for x_{n+1} that would be obtained by directly conditioning on $\boldsymbol{x_1}, \boldsymbol{x_2}$. Observations can therefore be introduced gradually using a region-growing approach on the model vertices.

The threshold used in the acquisition function is computed from the Mahalanobis distances of all previous query points $\boldsymbol{X_{n+q}^{t-1}}$. These points have already been labeled as inliers on the label map. The threshold therefore changes whenever more points are included to the query points. Their Mahalanobis distances are $\boldsymbol{d_M^t}$. Each distance is computed between the mean of the marginal PPD $\mu_p^t(\boldsymbol{x_i})$ at $\boldsymbol{x_i}$, its corresponding point on the target $\boldsymbol{y_i^t}$ obtained by minimizing the correspondence function $d_c(\boldsymbol{\mu_p^t}(\boldsymbol{x_i}), \boldsymbol{y_i^t}, \Gamma_T)$. The distance uses the marginal PPD covariance matrix $\Sigma_p^t(\boldsymbol{x_i})$ at $\boldsymbol{x_i}$, such that

$$d_M^t(\boldsymbol{x_i^t}) = \sqrt{(\boldsymbol{y_i^t} - \boldsymbol{\mu_p^t}(\boldsymbol{x_i}))^T \Sigma_p^t(\boldsymbol{x_i})(\boldsymbol{y_i^t} - \boldsymbol{\mu_p^t}(\boldsymbol{x_i}))}. \tag{4.7}$$

The Mahalanobis distances of all the query points are approximately lognormal distributed, discussed later in section 4.3.2. Taking the logarithm of these distances maps them to a Gaussian distribution with parameters μ^t, σ^t, from which a confidence interval (CI) can be used to obtain a classification threshold. An evaluation of the influence of the CI on the reconstruction and labeling results is found in section 5.2.2, based on which the 0.3 confidence interval is chosen. The upper bound of this CI about the mean is equivalent to the 65% percentile, which is used to compute the threshold:

$$\tau_{upper}^t = P_{0.65}(\mathcal{N}(\mu^t, \sigma^t)). \tag{4.8}$$

With the updated Mahalanobis distances $\boldsymbol{d_M^t}$ at hand, the acquisition function evaluates the neighboring vertices of the previous query points $\boldsymbol{X_{n+q}^{t-1}}$. The vertices to be evaluated are termed $\boldsymbol{X_a^t}$. The acquisition function is used to obtain $\boldsymbol{X^t}$, which contains the previous query points as well as those from $\boldsymbol{X_a^t}$ that satisfy the acquisition function:

$$J(\boldsymbol{x})|\boldsymbol{\theta^t} = \begin{cases} 1, & \text{if } log(d_M^t(\boldsymbol{x})) \leq \tau_{upper}^t \ \forall \boldsymbol{x} \in \boldsymbol{X_a^t} \\ 0, & \text{otherwise} \end{cases} \tag{4.9}$$

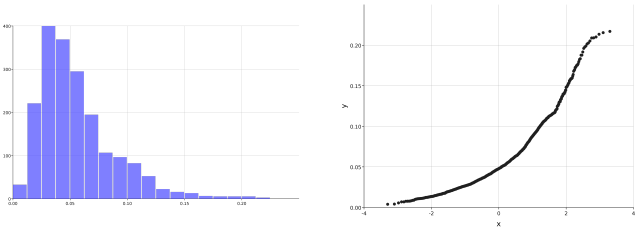


Figure 4.2: Histogram (left) and QQ plot (right) of inlier (all 2059 vertices) residuals in full-shape GP regression with ground-truth correspondences as input show a right-skewed distribution.

4.3 Assumptions

4.3.1 Gaussian Noise Model

The sequential GPMM algorithm assumes the Gaussian assumption on the noise in the inlier region is valid, enabling the use of the closed-form PPD. The histogram and QQ-plot in figure 4.2 show the residuals obtained from the ideal scenario with known correspondences. Instead of the straight line expected for a Gaussian distribution, the QQ-plot reveals a right-skewed distribution for the residuals. Despite the violation of the assumption, the deviation does not cause a reduction in the reconstruction quality in practice for healthy targets. Unlike the full target case, where all the reference vertices are used, the partial target case uses only those found in $\boldsymbol{X}^{\boldsymbol{t}}$.
Dependence on this assumption is common in robust inference. For example, this is the motivation behind RANSAC, where non-robust standard regression can be used in the sampled inlier region because of the assumption of sound statistical properties [1]. The proposed algorithm can also be compared to the standard pipeline provided in the GPMM registration pipeline [2]. The initial pipeline relies on the robust Huber loss function to filter outliers, making the closed form PPD solution valid. The proposed pipeline instead relies on a dynamic threshold estimated from remaining PPD variance and distance between mean and target.

4.3.2 Lognormal Distribution of Mahalanobis Distances

The Mahalanobis distances at the query points $\boldsymbol{X}^{\boldsymbol{t}}$ are calculated using equation 4.7. Figure 4.3 shows the histograms of these distances, obtained

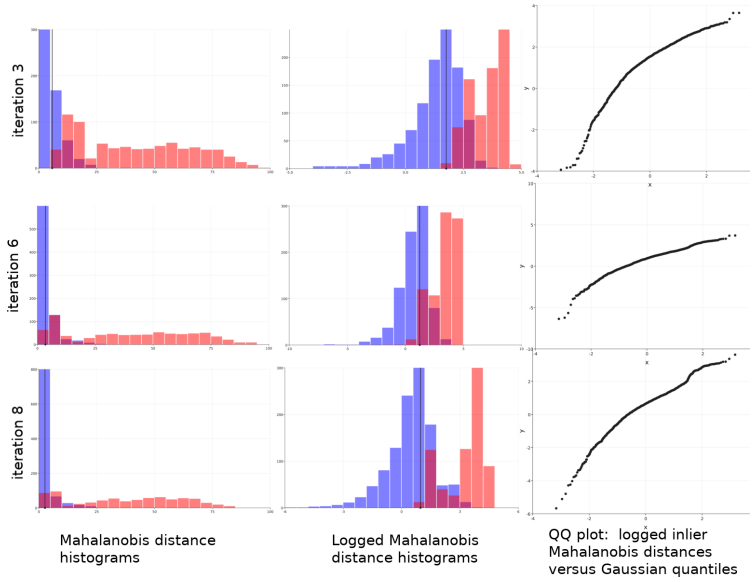


Figure 4.3: Mahalanobis distance histograms of the current inlier (blue) and outlier (red) obtained during sequential GPMM regression of a target with true inlier ratio = 0.66. First two rows (iteration steps 3 and 6): in red are domain points not yet evaluated. Third row (final iteration): in red are outliers detected as pathologies. Neighboring points of inliers with their distances below the threshold (vertical black line) are updated to inliers. The distribution of logged Mahalanobis distances in the second column is Gaussian distributed, confirmed by the QQ-plots in the third column. This figure is taken from [66].

during the sequential GPMM algorithm when fitting to a partial target. With every iteration, the number of query points $\boldsymbol{X}^{\boldsymbol{t}}$ shown in blue increases, because more reference vertices are included as query points as the algorithm proceeds. The remaining vertices in the GPMM reference are not yet used as query points and are shown in red. After convergence, the query points are labeled as inliers while the remaining model vertices are labeled as outliers. The plots show that the logarithm of distances tend to become less skewed and more similar to the expected distribution in figure 4.2 as more query points are used in inference. A confidence interval in the logarithm space is used to determine the threshold used in the acquisition function in equation 4.9.

The Mahalanobis distances are related to the likelihood of the observation

value under the marginal distribution obtained from the PPD at the observation. The likelihood under the probability density function is defined as

$$p(\boldsymbol{x_i^t}) = \frac{e^{d_M(\boldsymbol{x_i^t})^2}}{\sqrt{(2\pi)^k \left|\Sigma_p^t(\boldsymbol{x_i})\right|}}. \tag{4.10}$$

This implies that the squared Mahalanobis distance is proportional to the negative log-likelihood. To check how such data can be mapped to a Gaussian distribution, a box-cox transformation is applied to the distances to normalize them [67]. The box-cox transformation finds the transformation parameter λ that is applied to the data Y

$$T(Y) = \frac{y^\lambda - 1}{\lambda}, \tag{4.11}$$

such that the correlation coefficient between a Gaussian distribution and $T(Y)$ is maximized, computed from the values used to build a QQ-plot. When $\lambda = 0$, the logarithm transformation is applied. To reproduce the Mahalanobis distances obtained during a sequential GPMM threshold updating step, an example sample is created. The sample is created by sampling from a standard Gaussian distribution 1000 observations. The log-likelihoods of these observations are computed and normalized using equation 4.10 to obtain the Mahalanobis distances. The Mahalanobis distances are then passed through the box-cox test, with optimal transformation found at $\lambda = 0.42$, confirming that the data was right-skewed before the transformation because lambda is less than 0.5 [68]. This is almost proportional to a square-root transformation. The box-cox normality function that is maximized is shown in figure 4.4, in which the optimal λ line is plotted in red.

The data before and after the transformation is shown in the top row of figure 4.5. A logarithm transformation can be applied instead, with $\lambda = 0$, with minor changes to the function value in the normality plot. The transformed distribution has less correlation to the normal distribution, shown in the bottom row of the figure, but resembles the one obtained in practice in figure 4.3. The discrepancy could be due to the fact the residuals are only approximately Gaussian distributed to start with, even in the case of a healthy target reconstruction, as shown in figure 4.2 which reveals a skewed distribution for the residuals in a full GPMM regression. The skewness of the box-cox tranformed data is due to the standard box-cox estimator which is sensitive to noisy data. Nevertheless, central normality is achieved, where the transformed data look normal in the center with outliers in the extremes [68].

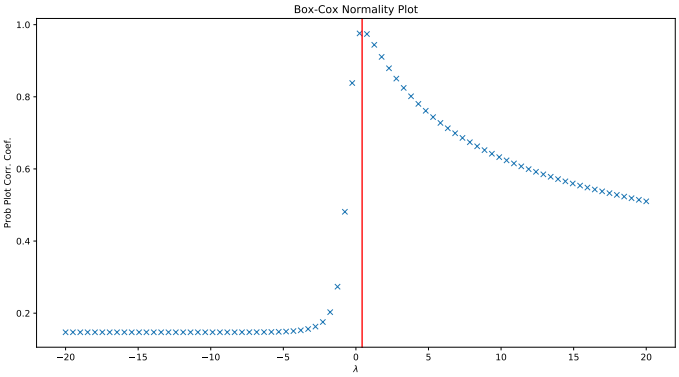


Figure 4.4: Box-cox normality plot, with the λ value that maximizes the correlation between the transformed data and the standard normal distribution plotted in red.

The transformation is therefore sufficient and it is not necessary to use more robust methods to acheive central normality [68].

4.4 Discussion: Comparison to robust GP inference

This section compares the proposed sequential GP regression approach to previous methods using the Neal dataset [6]. Observations generated by the Neal function $0.3 + 0.4x + 0.5\sin(2.7x) + \frac{1.1}{1+x^2}$ are corrupted with Gaussian noise as well as outlier noise to generate different outlier types. Seven types of outliers are taken from the literature [6].:

- rare outliers (5% outliers, outlier noise $\mathcal{N}(0, 1)$)
- fiducial outliers (15% outliers, outlier noise $\mathcal{N}(0, 1)$)
- abundant outliers (45% outliers, outlier noise $\mathcal{N}(0, 1)$)
- skewed outliers (15% outliers, outlier noise $\mathcal{N}(2, 1)$)
- extreme outliers (15% outliers, outlier noise $\mathcal{N}(0, 5)$)
- uniform outliers (30% outliers, outlier noise $\mathcal{U}(-3, 3)$)

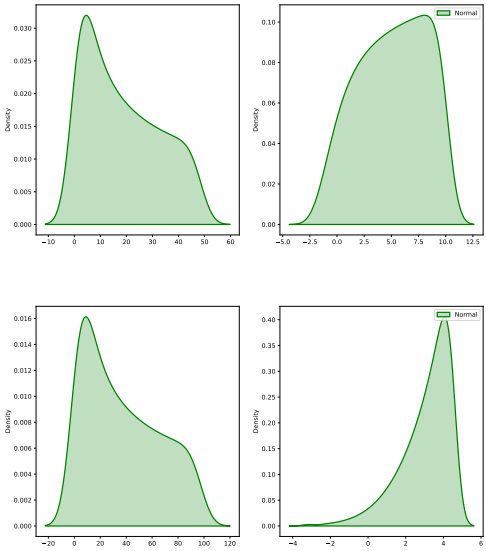


Figure 4.5: Top: Box-cox test applied to a sample of Mahalanobis distances computed from the log-likelihood values sampled from a standard Gaussian distribution. The calculation is based on the probability density function equation 4.10. Applying the box-cox transformation with $\lambda = 0.42$ provides the closest mapping of the distances from the initial distribution (left) to the normal distribution (right). Bottom: The logarithm transformation maps the same sample of Mahalanobis distances to the distribution on the right, which is closer to the observed behavior in figure 4.3.

- Student distribution with 3 degrees of freedom outliers (100% outliers, outlier noise t_3 distributed)

Two more types are introduced to represent pathologies:

- holes, where a group of neighboring observations are removed (10% outliers, outlier observations are removed from the provided training set)
- bumps, where a group of neighboring observation values are replaced with the GP mean values (10% outliers, outlier observation values replaced with the GP zero-mean)

For the experiments, 500 observation locations are randomly sampled from the x-axis and used to generate the observations with noise using the Neal function and the outlier type. From those observations, the 15 observations (3%) with the smallest residuals to the ground truth Neal function are chosen as initial landmarks.
Robust Gaussian Process inference can be split into two groups. The first replaces the Gaussian noise term with a heavy-tailed distribution that is more appropriate for modeling outliers, such as the Laplacian or the Student-t distribution. The advantage of using the heavy-tailed likelihood is that the outlier filtering step is no longer necessary. The disadvantage is that the closed-form predictive posterior distribution is no longer valid, and approximations of the posterior distribution become necessary. To perform approximate inference of the GP parameters and obtain the PPD, the Expectation Propagation algorithm is used. The following methods are chosen to represent this group:

- T-likelihood GP (tlikeGP): the Gaussian likelihood term is replaced with the heavy-tailed student-t distribution with 5 degrees of freedom [6].
- Bayesian Optimization with Outliers (bowoGP): regression is split into two steps. In the first part, GP regression with a student-t distribution likelihood term is performed on a chosen subset of the observations. Using the robust regression result, the observation residual distribution can be learned. The upper and lower 5^{th} percentile observations are labeled as outliers, and the remaining observations are used to obtain a more accurate regression result by using a Gaussian likelihood error term. This is repeated iteratively for a specific number of iterations or until the expected ratio of outliers has been reached [1].

The second group aims to keep the Gaussian noise term and analytical predictive posterior distribution formulation. With this strategy, the outliers must be separately detected and removed from the observation set to enable valid PPD inference. The following iterative methods are chosen to represent this group:

- Huber GP (hubGP): the residuals are passed through a robust Huber loss. The function is quadratic for residuals below a threshold δ and linear for those above δ. This is followed by a classification step which filters out corrected residuals that are above a threshold. This is similar to the distance correction step in the initial GPMM pipeline [2].
- Iterative Trimming GP (itGP) [6]: GP regression is performed iteratively. For the initialization step, all the observations are used for regression. The normalized (Mahalanobis) distances between the observations and the PPD mean are calculated and sorted, after which α observations having the smallest residuals are chosen and used in the regression of the next step.

The four methods are compared to the proposed sequential GP algorithm (seqGP) and traditional GP inference (tradGP) in optimization and regression experiments. The average over 10 differently initialized runs is used to obtain the mean squared error (MSE) and runtimes. In these experiments, the number of landmarks provided for initialization is 10, representing 2% of the total observations.
The top row in figure 4.6 and figure 4.7 show an inference result for the different outlier cases. The inlier observations are in blue, while the outlier observations are in red. The ground truth function from which the observations were sampled is in dashed black. The bottom row of each figure shows the mean and standard deviation of inference results from 10 different iterations. In each iteration, a different sample of observations is taken from the ground truth function to be used as input and for model initialization. The seqGP method shows competitive results by providing reconstructions that are as good as other methods. seqGP also shows promising results in the smooth outlier cases, which were introduced to represent clustered pathology-like outliers, as seen in figure 4.7. Similar results are obtained for the sparse outlier types, so only the extreme outliers case is shown in this chapter and the remaining cases are presented in appendix B. Similarly, for the smooth outlier types, the bump case is shown in this chapter while the holes case can be found in appendix B.

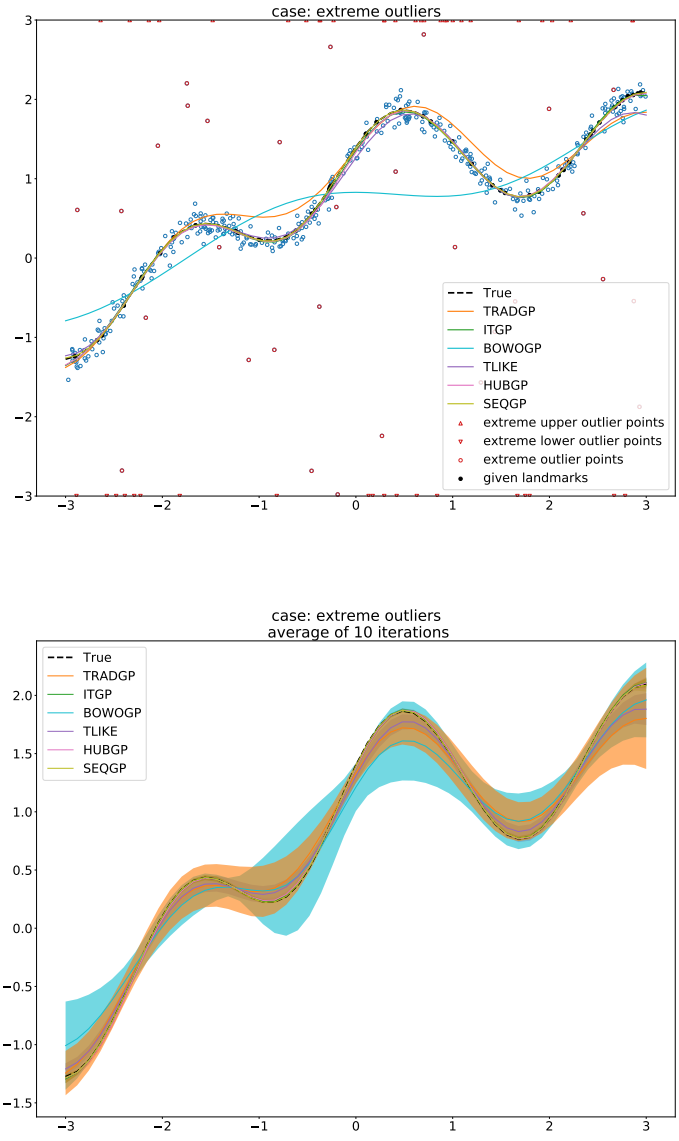


Figure 4.6: Case: Extreme outliers. Top: mean of the inferred GP obtained from the different robust GP inference methods. Bottom: Average mean and standard deviation of the inferred GP mean from 10 different iterations.

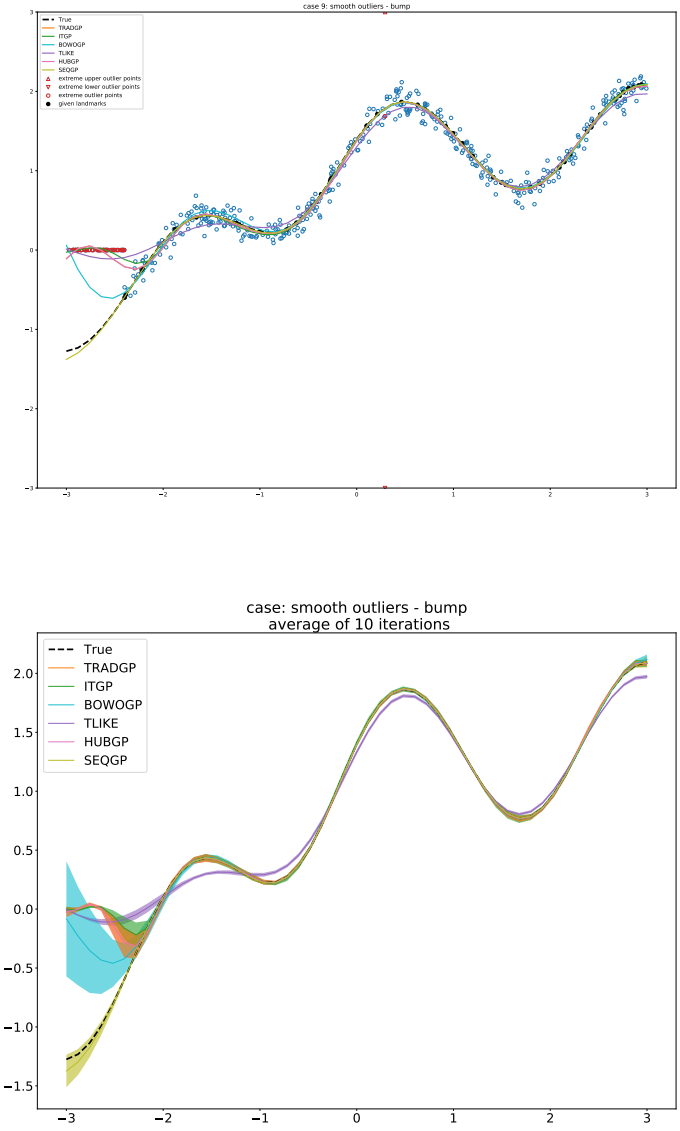
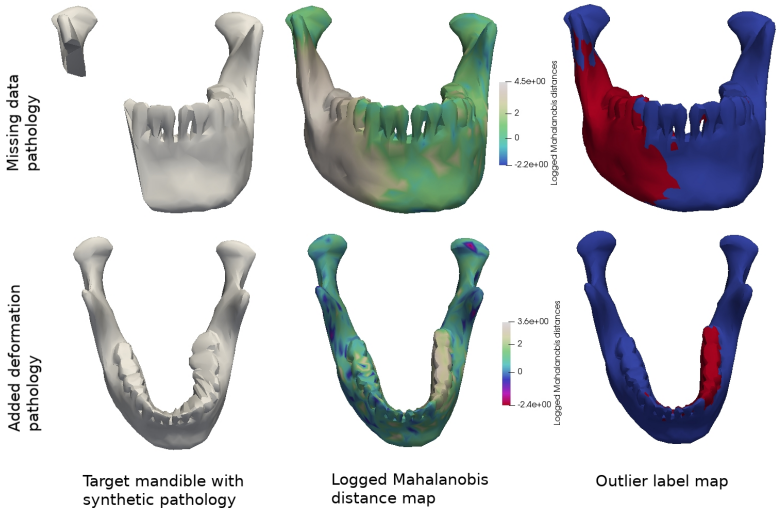


Figure 4.7: Case: Smooth bump outliers. Top: mean of the inferred GP obtained from the different robust GP inference methods. Bottom: Average mean and standard deviation of the inferred GP mean from 10 different iterations.

Chapter 5

Evaluation

A regression algorithm is considered robust if its regressed output is not influenced by the outliers. In other words, a reconstruction in the inlier region should be similar to the reconstruction obtained if the target did not have any outliers. This section evaluates how robust the proposed solutions from chapters 3 and 4 are.
All evaluations are performed on a synthetic dataset, built by introducing pathologies that influence a subset of the reference shape vertices. In this way, ground truth reconstructions as well as label maps are available for evaluation. The output is considered robust if it produces a shape reconstruction with low average and Hausdorff distances to the ground truth shape and accurate label predictions compared to the ground truth label map. This is quantified using the F1 score, which provides a summary of the confusion matrix in cases of positive and negative class imbalance. It is computed as the harmonic average of precision and recall. Examples are shown in figure 5.1. The first column shows the target mandible with the introduced pathologies. The pathologies take the form of deformations out of the shape model space introduced to a subset of the target vertices, or removed vertices that represent missing data pathologies. The second column shows the Mahalanobis distances that are used to obtain the threshold. The third column shows an overlay of the binary label map and the target. The pathology prediction is indicated in red, while the blue region shows the predicted inlier region used to obtain the PPD.
The proposed methods are first compared to previous robust shape modeling strategies from the literature. This is followed by a breakdown point analysis and sensitivity analysis of the proposed methods.



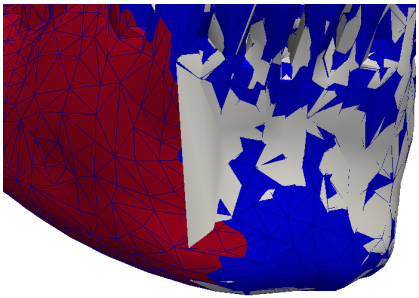


Figure 5.1: Top: Example synthetic pathologies that have been introduced to a sample from the mandible SSM. The second column shows the final Mahalanobis distances, and the third column shows an overlay of the target and final label map with predicted pathology regions in red. Bottom: close-up on the edge of the target with missing data, overlayed with a meshed surface of the reconstruction with label map. The overestimated pathology region is due to the effects of the underlying meshing of the fit on the visualization. The fit does not have vertices on the border of the pathology, but instead interpolates between the inlier and outlier vertices to obtain the surface color.

5.1 Comparison to Previous Approaches

The following robust surface reconstruction approaches are compared:

1. **RANSAC** [69] falls into the reconstruction-based approaches with region growing explained earlier in section 2.1.1. The implementation is based on [41] and initialized by setting a random 30% of model vertices as inlier landmarks, from which the closed-form PPD is calculated. The consensus set is found based on the PPD: points in the reconstruction with Mahalanobis distance less than 3.0. If the consensus set consists of more than 60% of the model vertices, the labels are updated. The PPD is updated using the new label map. Fitting is repeated, each time with a different random sample of inlier vertices. The reconstruction with the lowest residual errors is finally chosen as the fit.

2. **Coherent Point Drift (CPD)** [38] falls into the probabilistic approaches explained earlier in section 2.1.2 and based on the implementation from a thesis in the group[1]. CPD formulates the registration problem as a mixture model density estimation problem, where one point set represents model means and the other samples. Since the true pathology ratio in the synthetic experiments is known, the weight vertices w are set as the true outlier ratio. After convergence, the label map can be obtained from the sparse correspondence matrix. In real applications, w is not known, requiring different values to be tested or other approximate methods.

3. **Nonrigid ICP (NICP)** [32] falls into the reconstruction-based approaches over the full reference explained earlier in section 2.1.1. NICP is a robust extension of ICP with optimal update steps. The implementation is provided in the same repository as the CPD one. The regularization hyperparameter for template stiffness is initialized to $\alpha = 512$ and halved every 2 iterations until it reaches 0.5. After convergence, the label map can be obtained from the vertex weights learned by the algorithm [32].

4. **Forward-search SSM with informed proposal (FSSM+IP)** FSSM+IP is the EM-like algorithm approach explained in chapter 3

[1] Code for template registration with scala: https://github.com/madsendennis/template-registration-with-scala

that iteratively approximates the PPD by sampling and region growing. A probabilistic implementation of the Iterative Closest Points algorithm [70] is used to make informed proposals and speed up fitting.

5. **Sequential GPMM (SeqGPMM)** SeqGPMM is the closed-form PPD approach explained in chapter 4.

The average distance (best at 0mm), Hausdorff distance (best at 0mm), and F1 score (best at 1.0) are used to summarize the results. Figure 5.2 reveals that the two proposed methods are least influenced by the ratio of outliers and show better reconstruction and detection results across the three metrics.

5.2 Discussion

The sensitivity and breakdown point analyses are provided in this section for the sequential GPMM algorithm. The hyperparameter influence on the forward SSM algorithm is similar to that on the sequential GPMM. The forward SSM plots are therefore only shown in Appendix C for completion.

5.2.1 Breakdown Point Analysis

The breakdown point indicates the signal-to-noise ratio ranges within which the algorithm is capable of producing a robust result [8]. This is determined by evaluating the algorithm performance at different outlier ratios. The outlier percentages used for testing are: 0.1%, 1%, 2%, 5%, 10%, 20%, 25%, 33%, 50%, 75%, 83%. Figure 5.3 shows some examples of these missing percentages as well as reconstructions and label maps.
In terms of reconstruction, the method fails for ratios greater than 50% of the target size when the bounding box of the pathological target no longer covers the ground truth bounding box of its healthy shape. This is revealed by the increase in the AD and HD values in figure 5.4. This is also shown in the qualitative results of the figure 5.3. The fourth row shows a successful reconstruction case with 50% outliers, while the last row shows a failed case for that same percentage. In the successful case, the left condyle is available despite the missing data and can guide the reconstruction of the outlier region.
On the other hand, small outlier percentages cause the algorithm to succeed in the reconstruction task but fail in the detection task. This is shown by the low F1 score in figure 5.4 when outliers influence less than 10% of the target. As the signal-to-noise ratio increases, the pathologies become smaller and

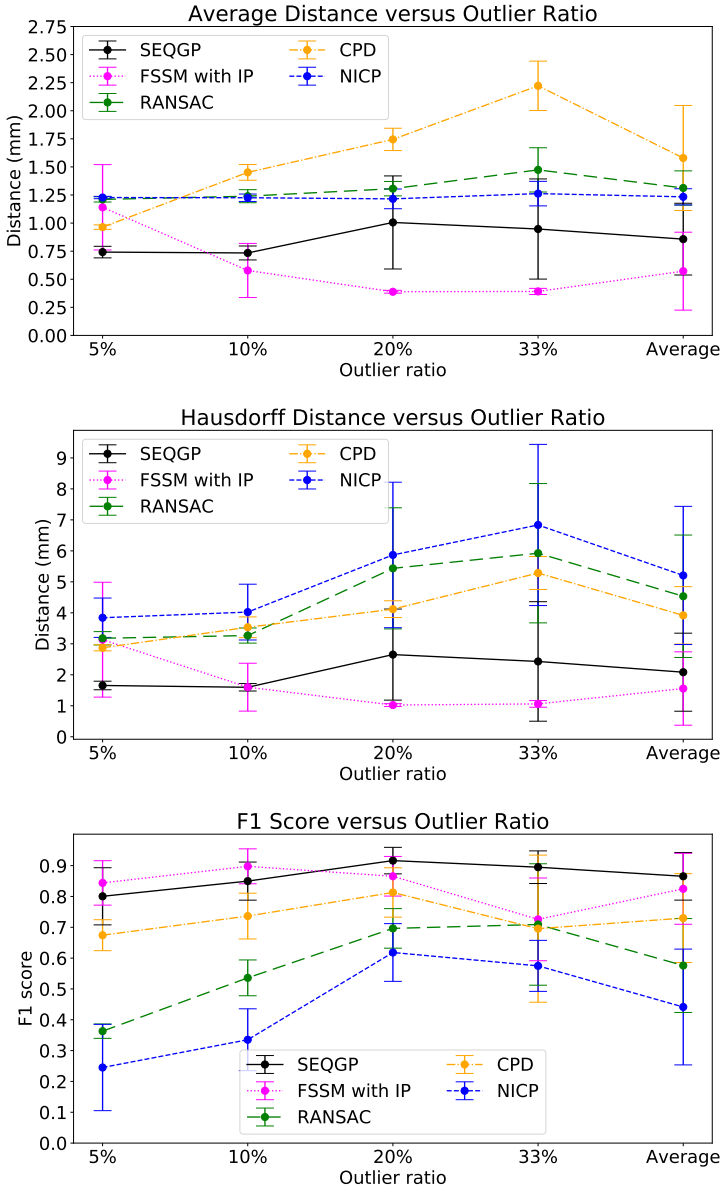


Figure 5.2: The proposed approaches (SEQGP and FSSM with IP) perform better than the other reconstruction approaches across different outlier percentages. The average distance (top) and Hausdorff distance (middle) are best at 0 mm, while the F1 score (bottom) is best at 1.0.

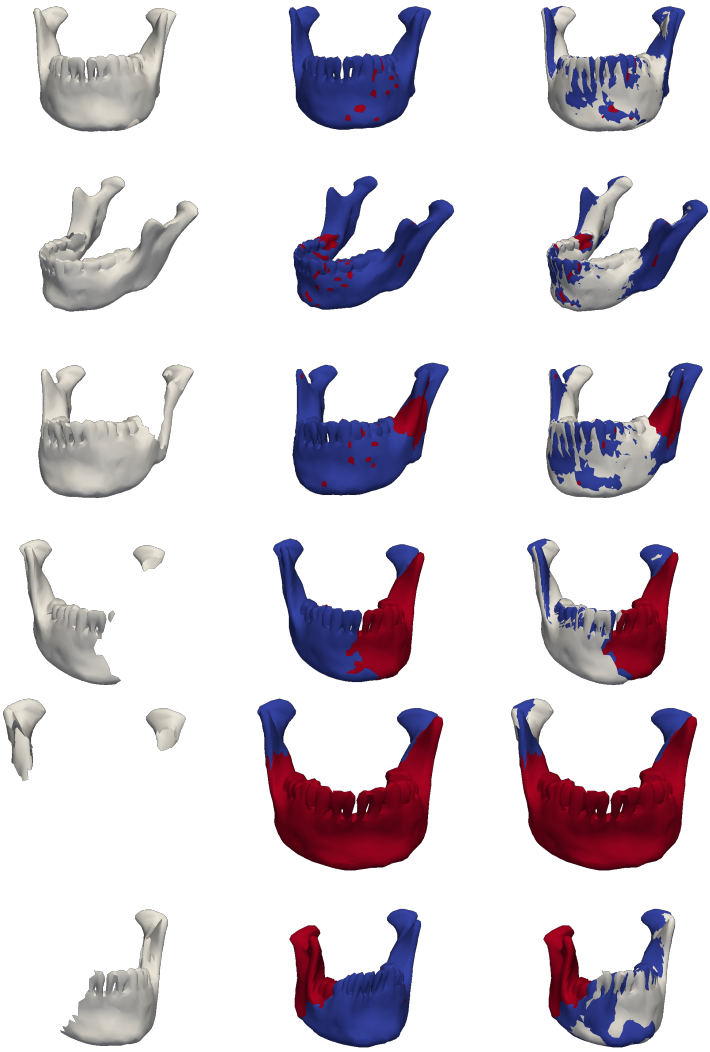


Figure 5.3: Qualitative results on the synthetic outliers introduced to the mandible SSM sample, for the following percentages of the breakdown analysis: 0.1%, 1%, 10%, 50%, 83%. The last row is an example of a failed reconstruction but successful outlier detection which occurs with a 50% outlier case, which occurs because the missing part of the mandible causes a loss of information about the location of the right condyle.

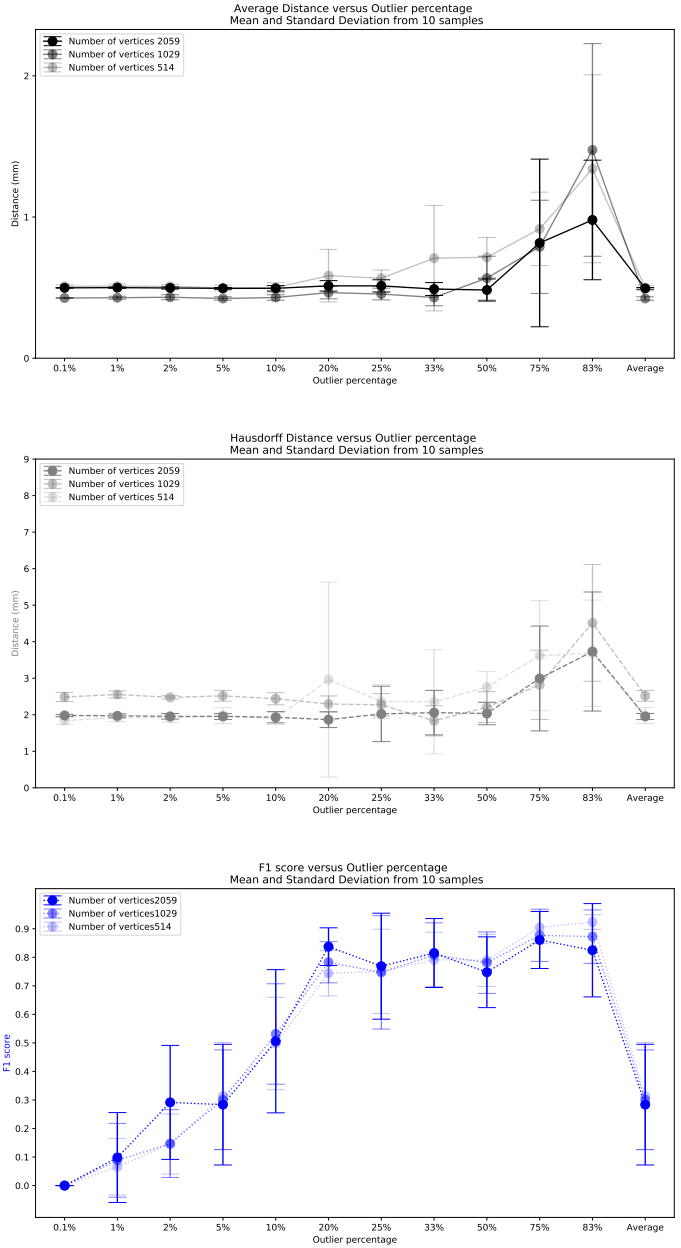


Figure 5.4: Breakdown point analysis for sequential GPMMM. There is a decrease in reconstruction scores but an improvement in the detection score as the outlier percentage increases. The color shading shows that the model reference mesh density does not influence the results.

can no longer be detected, which is why the F1 score drops despite the good reconstruction performance. With 0.1%, only 2 vertices on the target have been given pathological deformations. The correspondence function fails to map the correct point on the target to the reference, and the provided correspondence match always provides a distance smaller than the classification threshold. As a result, there will be false healthy labels on the reference in the inferred label map. An example is the first row in figure 5.3. The small hole on the bottom left of the chin is not detected by the algorithm. This problem is independent of the reference mesh density, which can be seen by comparing the F1 scores at the same outlier percentage for different reference mesh resolutions in figure 5.4. Some practical applications where this problem might come up are high frequency pathologies and pathologies that lie within the healthy shape range. When the signal-to-noise ratio is low, the percentage of outliers is high. With 83% outliers such as in the row before last in 5.3, 1705 of the 2000 landmarks are pathological. The F1 score remains high despite the drop in reconstruction quality, because the initial landmark alignment is sufficient to generate a pathology region prediction that is accurate enough.

5.2.2 Evaluation of Hyperparameter Influence

The hyperparameters that can influence performance are: model flexibility, quantified by the model rank, and initialization, quantified by the number of landmarks provided as inliers. Higher flexibility and more landmarks improve the reconstruction distance and detection scores. This can be seen in figures 5.5 and 5.6. This behavior is independent from the outlier percentage in the target, seen by comparing the 5% outliers case with the 33% outliers case. Model flexibility can be evaluated through the training set size instead of the model rank. Figure 5.7 shows best performance for the augmented model, with decreasing reconstruction and detection qualities as the size of training set drops.
Furthermore, the influence of the reference mesh density is evaluated. Different models are built by changing the reference mesh density. The breakdown analysis figure 5.4 indicates with the shading of the plots the influence of model density on the reconstruction and detection scores. The Hausdorff distance is affected by the number of vertices in the model and shows best performance when the SSM density is largest. This is because the metric is computed by choosing the maximum distance computed between the model reconstruction vertices to the target and vice versa, and as such can be influenced by differences in the model and target densities. This effect is

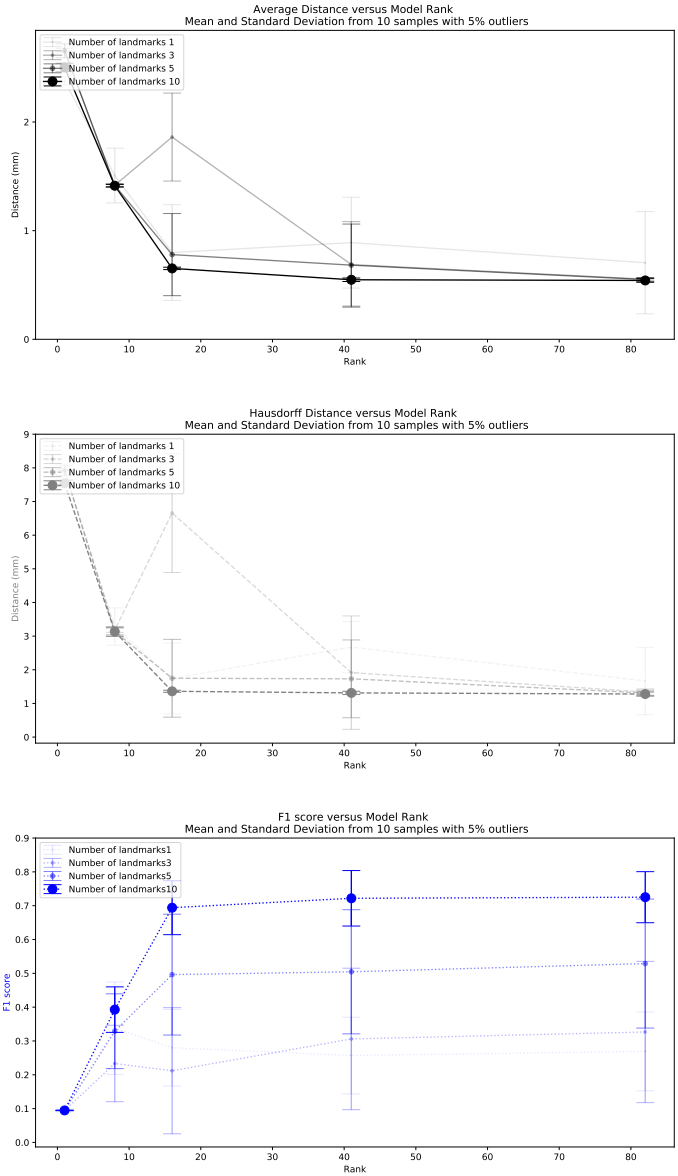


Figure 5.5: Performance evaluation on targets with an outlier percentage of 5% shows an improvement in the reconstruction and detection scores as the model flexibility increases, quantified by the rank. The number of landmarks used for initialization does not influence the reconstruction metrics once the model is flexible enough, and only the detection score improves when the number of initial landmarks is larger.

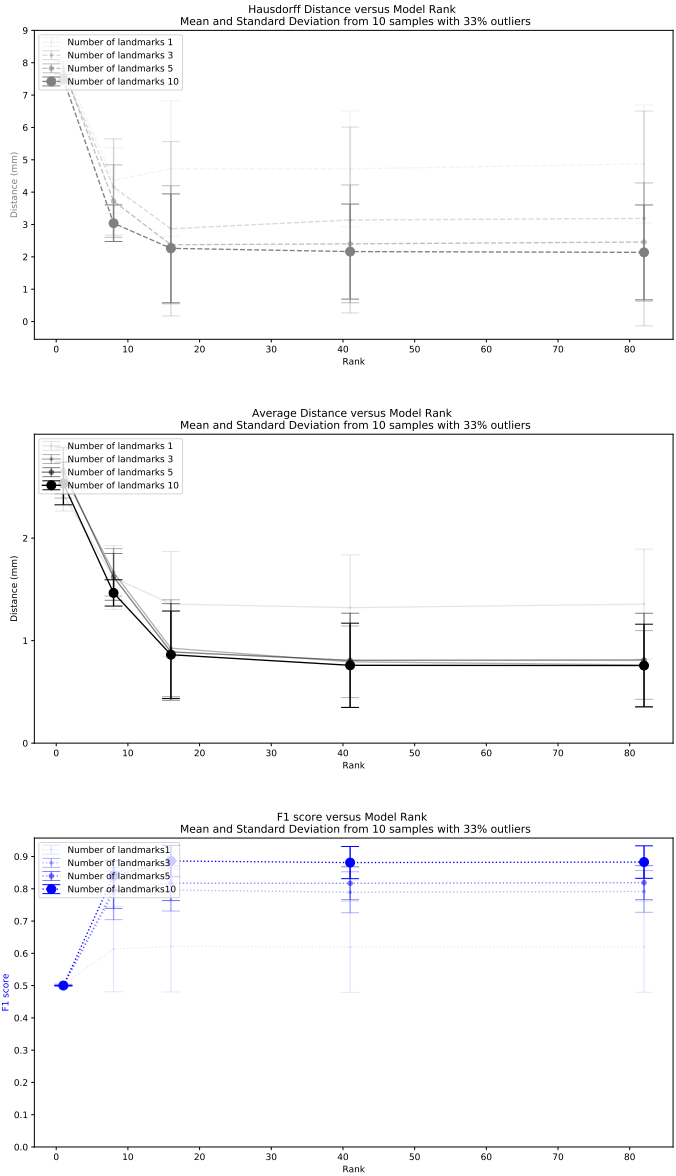


Figure 5.6: Similar to the 5% outlier case, the performance evaluation on targets with an outlier percentage of 33% shows an improvement in the reconstruction and detection scores as the model flexibility increases, quantified by the rank. However, the opposite behavior is observed when considering the effect of the initial landmark number: the number of landmarks used for initialization influences the detection metric less than it affects the reconstruction

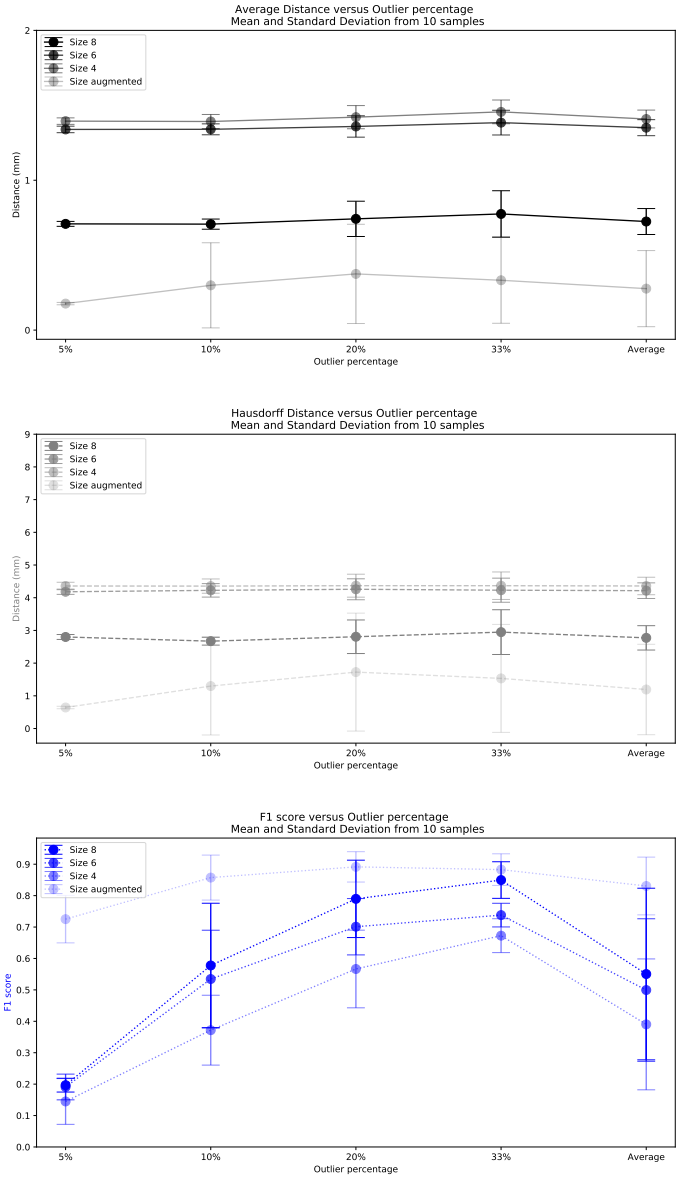


Figure 5.7: Performance evaluation across different training set sizes used for model building reveals that higher model flexibility, quantified by the number of training examples, improves the reconstruction and detection scores.

mitigated once the average distance is used as metric. The density does not play a role in the F1 scores either.
Finally, the confidence interval (CI) and the observation uncertainty used for the landmarks are evaluated. The CI is applied to the logarithm of the Mahalanobis distances when updating the threshold, as seen in equations 4.8 and 4.9. The landmark uncertainty influences the remaining covariance in the PPD, appearing in the form of σ in equations 1.4 and 1.5. In practice, it determines how accurately the PPD mean reconstructs the observations. Higher uncertainty allows the PPD to maintain flexibility at the cost of less accurate reconstruction of the landmarks. To start with, the average distance plots in figure 5.8 reveals that the CI does not influence the final reconstruction accuracy for low landmark uncertainty levels. However, as the landmark uncertainty increases from the top to the bottom row, lower CI values give lower reconstruction distances. This can be viewed as a trade-off between the classification confidence and the reconstruction confidence. For low reconstruction confidence levels (high landmark uncertainty), more conservative CI values make it harder to falsely label a novel observation as inlier. This is confirmed by looking at the last row in figure 5.10, in which the F1 score performs better for lower CI levels. A safe CI of 0.3 is chosen with a landmark uncertainty of 1.0 for the applications in chapter 6.

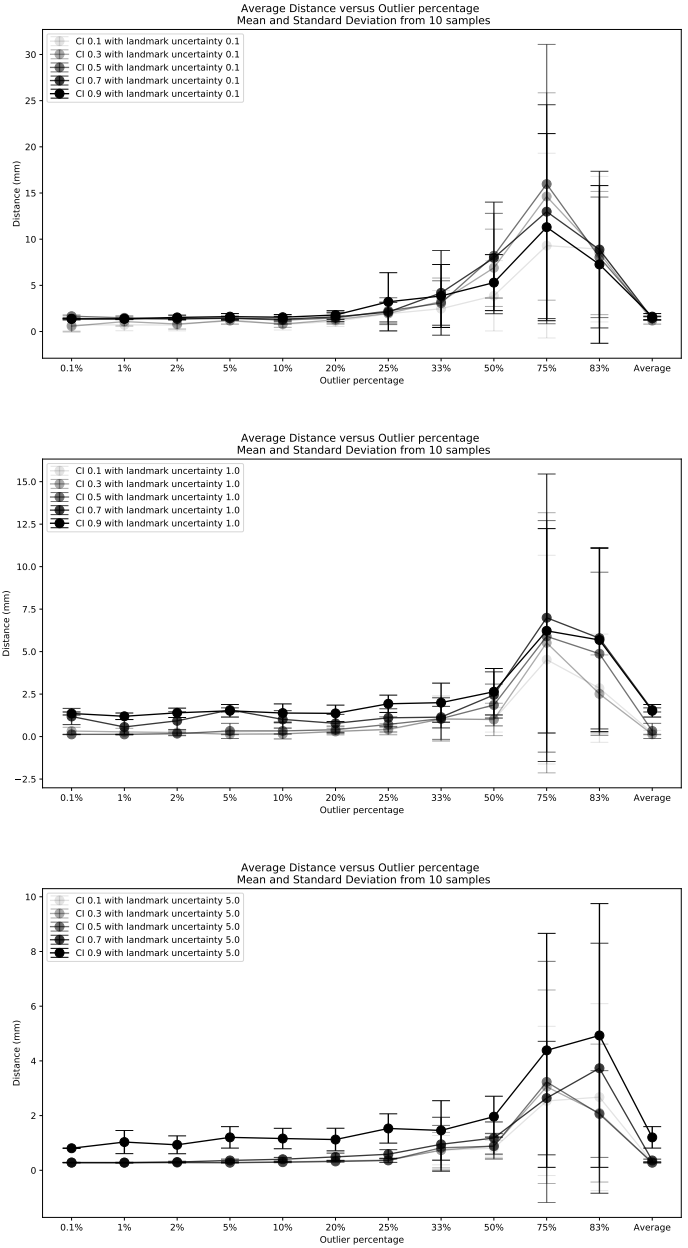


Figure 5.8: The chosen confidence interval does not influence the average distance reconstruction scores. However, larger landmark uncertainty improves the reconstruction average distance for large outlier percentages, seen by comparing the top, middle and bottom plots.

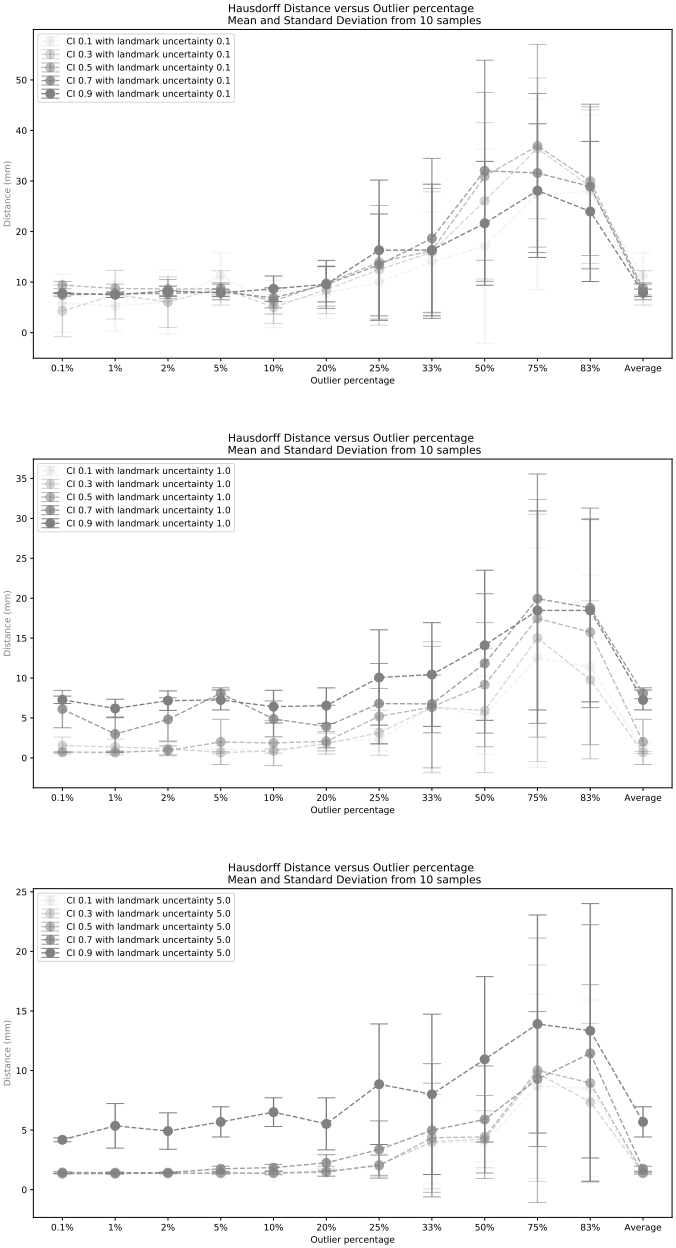


Figure 5.9: Similar to the average distance metric plots above, the chosen confidence interval does not influence the Hausdorff distance as much as the landmark uncertainty does. Larger landmark uncertainty improves the Hausdorff distance results, seen by comparing the top, middle and bottom plots.

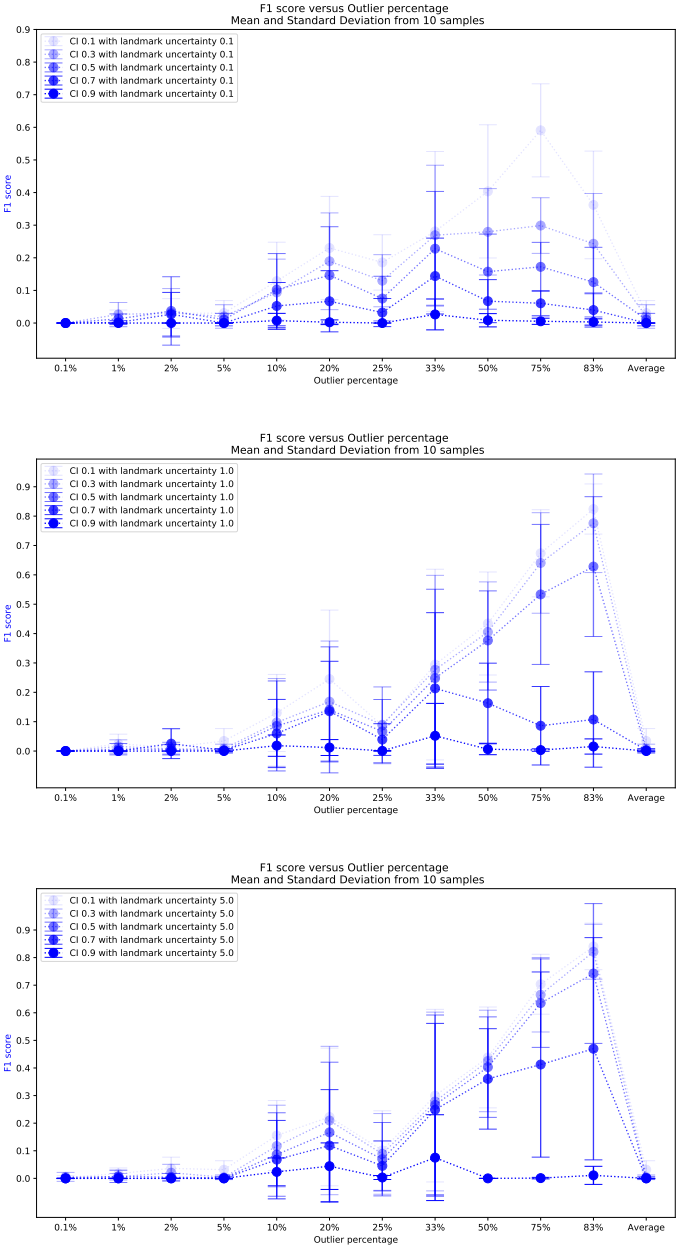


Figure 5.10: The chosen confidence interval influences the resulting detection scores. Smaller confidence intervals are more conservative, which explains the higher F1 scores of the plots in lighter blue. The landmark uncertainty has smaller influence on the detecion scores.

Chapter 6

Applications

Because the proposed methods are based on outlier detection, it is possible to directly apply them to different targets to perform pathology detection and shape reconstruction. The only requirement is a GPMM describing the shape family and some landmarks for alignment and initialization of the inlier regions. The generalization ability of sequential GPMM is shown in this section on various targets with different pathologies that highlight clinical relevance. The figures share the same color legend: the pathological target is shown in orange, the reconstructed mesh with label map overlay is shown in red (outliers) and blue (inliers), and the healthy shape of the pathological area of the target is shown in green when available.

6.1 AutoImplant2020 Challenge

The AutoImplant challenge, first presented at MICCAI 2020, aims to automate cranium implant design [71]. Craniums with holes are provided, and the goal is to provide an accurate reconstruction of the missing part. This challenge is an example of shape reconstruction with pathologies, where the pathologies in this case are the missing parts.
Sequential GPMM starts with a full skull GPMM. Only the cranium is used for the reconstruction. The label map output is used to indicate the part on the cranium that is missing. This region is extracted from the reconstruction as the implant. Sequential GPMM is tested on the 99 provided craniums, giving an average of 1.99 mm for average distance, 6.42 mm for Hausdorff distance, and 0.73 for dice score, all in the implant region. Two reconstruction and detection results are provided below as examples. The first, figure 6.1, is a

successful case, with 3.38 mm, 1.26 mm and 0.82 for HD, AD and F1 scores in the implant region. The results are comparable to the leading scores on the leaderboard, which have 3.96 and 0.94 HD and F1 scores. The second is figure 6.2, with 8.37 mm, 3.03 mm and 0.62 HD, AD and F1 scores. The pathology detection labels are correct, and they guide the reconstruction in the healthy region, giving the low inlier region errors in blue. However, the reconstruction of the missing region, which represents the implant, can be improved. This is shown with the underestimated implant thickness in the image slices. Outlier region inference can be improved by using a case-specific expected loss, explained in the future work section 7.2. The loss in this case could encourage a smooth surface transition from inlier to outlier regions. Because sequential GPMM does not optimize the reconstruction in the outlier region, it can reliably be used to obtain the map and an accurate reconstruction of the inlier region. To improve the accuracy of the missing region reconstruction, further reconstruction can be performed using the sequential GPMM reconstruction and label map for initialization [70].

6.2 KITS2019 Challenge

The KITS challenge was first presented at MICCAI 2019 with the goal of automating kidney tumor detection [72]. The challenge provides kidney images with various tumors. Each image contains the left and right kidneys, but not both of them are pathological. Multiple tumors can appear on the same kidney, they can be surface tumors or inner-organ tumors, and they have can have different sizes and shapes. This challenge therefore presents a good example of additional data outliers, where pathologies are caused by unlikely deformations under a shape family.

To run sequential GPMM, a kidney shape model is first built from the healthy kidneys in the dataset. Only 3 landmarks are used for alignment and initialization. Sequential GPMM is tested on the 109 pathological kidneys. The example shown in figure 6.3 has 0.86 and 0.81 dice scores for kidney and tumor, while that in figure 6.4 has dice scores of 0.85 and 0.32 respectively. The average scores are 0.65 and 0.62 for average dice scores of the reconstructed kidney and tumor regions, which is significantly lower than the leaderboard scores of over 0.9 and 0.85 respectively. There are various reasons behind the gap between proposed method and the challenge results. To begin with, the label map is defined on the reference mesh of the shape mesh in the sequential GPMM algorithm. To obtain a label map on the target, the reference label map is projected onto the target vertices according to the

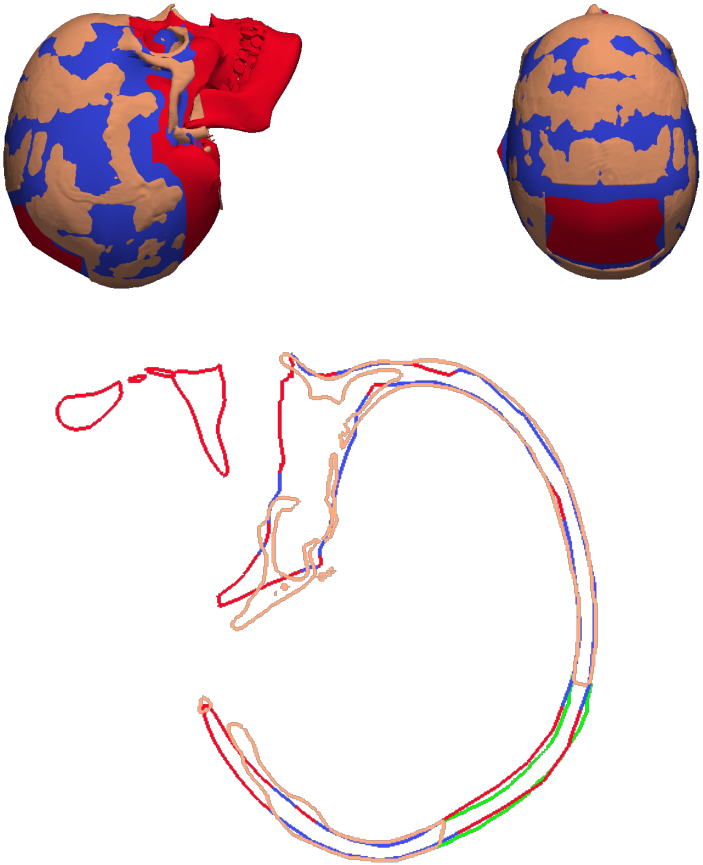


Figure 6.1: Example from the AutoImplant2020 cranium reconstruction challenge, with: target in orange, reconstruction with label map in blue and red, and ground truth implant in green.

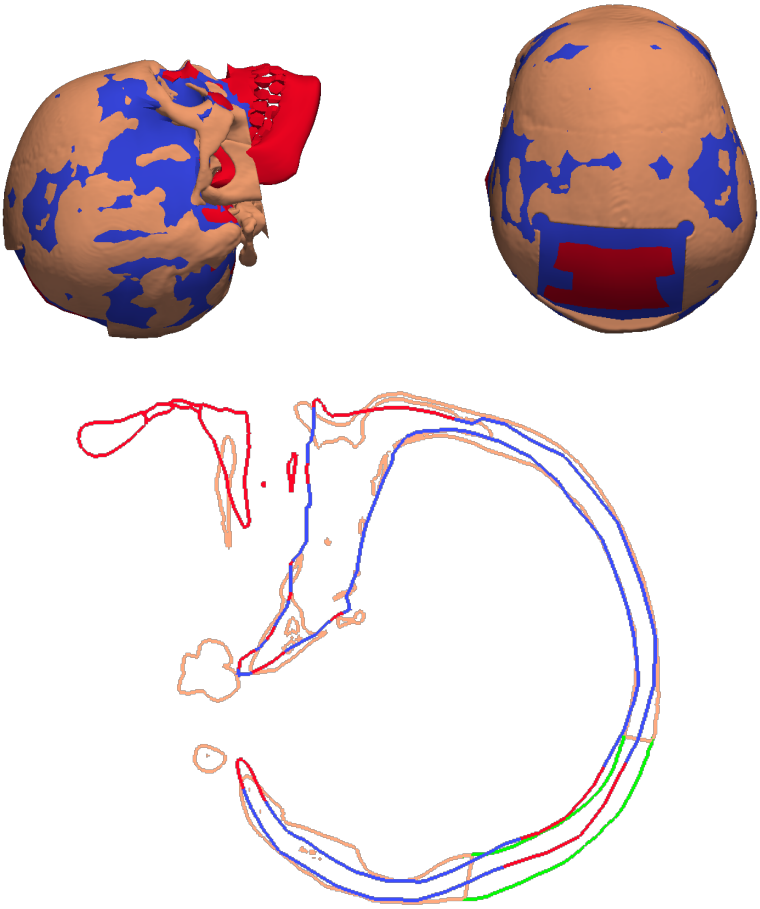


Figure 6.2: Example from the AutoImplant2020 cranium reconstruction challenge, with: target in orange, reconstruction with label map in blue and red, and ground truth implant in green.

correspondence function. This implies that points with correspondence will be labeled as inliers, while those without will remain labeled as outliers. This causes differences in the obtained label map compared to the leaderbard methods, which directly perform pathology segmentation in the image domain. Second, the shape model used in the proposed method is limited to surface deformations; in other words, inner volume tumors in the kidneys cannot be detected because there model reference is a hollow mesh. This results in a decrease in the dice score compared to the volumetric image methods. Different extensions that can be included in the sequential GPMM algorithm to enable inner-volume pathology detection are discussed in chapter 7. Finally, the healthy shape model of the kidney has high flexibility. Although this enables the model is able to generalize well to novel targets, it also implies that the surface bump deformations are included in the space of healthy deformations. Therefore, the Mahalanobis distance threshold will fail to exclude some vertices associated with tumor pathologies, because these vertices lie within the threshold.

6.3 Shape2015 Statistical Shape Model Challenge

The Shape 2015 challenge consists of the reconstruction of 10 partial livers. The partial targets as well as training set of full livers are available on the Virtual Skeleton Database [73]. A liver model is built from 20 of the registered full livers. After an initial rigid alignment based on landmarks clicked on physiologically relevant locations by the user, the reconstructions from the sequential GPMM and the traditional GP regression are obtained. The full shapes of the 10 partial targets were not available on the online database, neither was the automatic evaluation system still accessible at the time of this thesis. Therefore, the reconstructions are evaluated only in the inlier regions using the average distance. The violin plots in figure 6.5 compares the average and Hausdorff distances obtained using different reconstruction methods for the different partial targets. The first is the proposed sequential GPMM reconstruction, the second relies on the traditional GPMM fitting algorithm with correspondences established starting from the target vertices, and the third uses a model first conditioned on the landmarks. The proposed sequential GPMM algorithm has smaller average distances between the partial target and the reconstructions, which is an expected result for the robust reconstruction algorithm. Nevertheless, the two algorithms perform similarly well, because the rigid alignment performed to

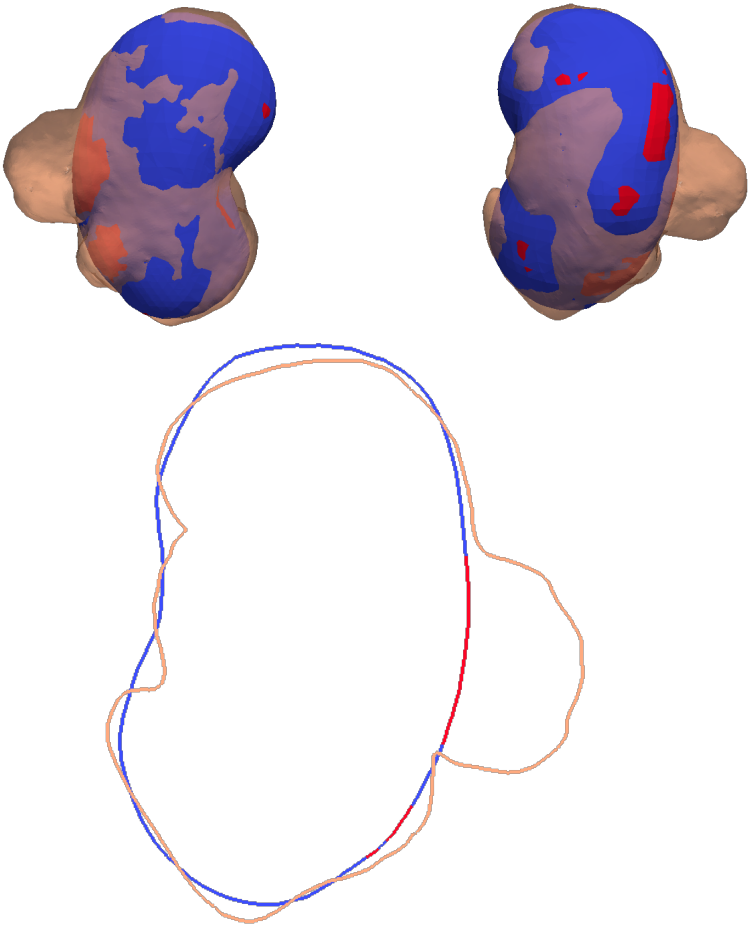


Figure 6.3: Example from the KITS19 tumor detection challenge, with: target in orange, reconstruction with label map in blue and red.

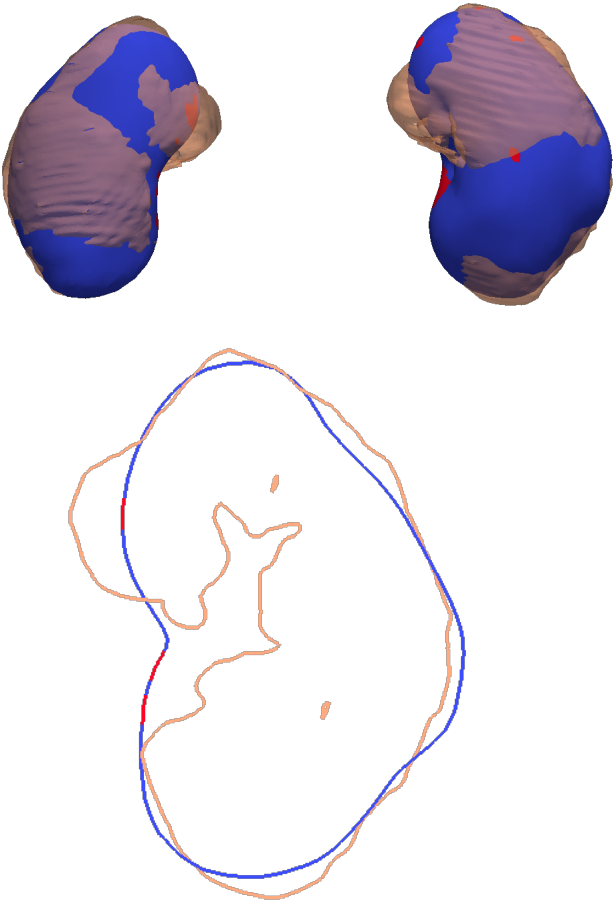


Figure 6.4: Example from the KITS19 tumor detection challenge, with: target in orange, reconstruction with label map in blue and red.

initialize the fitting provides sufficient information for the reconstruction; i.e. after finding the corresponding point for every vertex on the target to a vertex on the model, the model vertices which have not been matched indicate outlier regions. The visualization in figure 6.7 shows an example reconstruction overlayed with the predicted label map (red and dark blue), the reconstruction obtained with traditional GPMM regression without a label map but with a target-to-model correspondence function (light blue), and the target (orange). The target in this case is missing a part of its surface on the posterior side of the liver. Another example is shown in figure 6.8, where the target is missing the left lobe. The two targets are visualized separately in figure 6.6.

6.4 Clinical Examples

This section shows results on clinical data. The ground-truth shapes and label maps are not available, which is why qualitative results are shown below for the femur and mandible examples [2] and only reconstruction distances to the inlier region can be used as evaluation metric.

6.4.1 Forensics dataset

The next examples are taken from the forensics dataset. The forensics dataset consists of 86 skulls. The pathologies that are found in this dataset are in the mandible, due to missing teeth and teeth artifacts generated by the imaging process. Three example targets which will be discussed in more detail are shown in figure 6.9.
The results are summarized in figure 6.10 for the proposed method (sequential GPMM), the standard GP regression method which uses all the vertices on the model reference (All), and the standard GP regression method which conditions the model only on the provided input landmarks (Landmarks). The distances are computed only in the inlier region. The region is defined according to the label map generated using the Sequential GPMM, because the other two approaches do not rely on a label map. The results show that excluding the outliers from the observations to be used in GP regression does improve the reconstruction results in the healthy regions.

[2] The clinical examples have been provided by: Institute of Forensics at the University of Zurich for the skulls, Uniklinik RWTH Aachen and Queensland University of Technology for the femur.

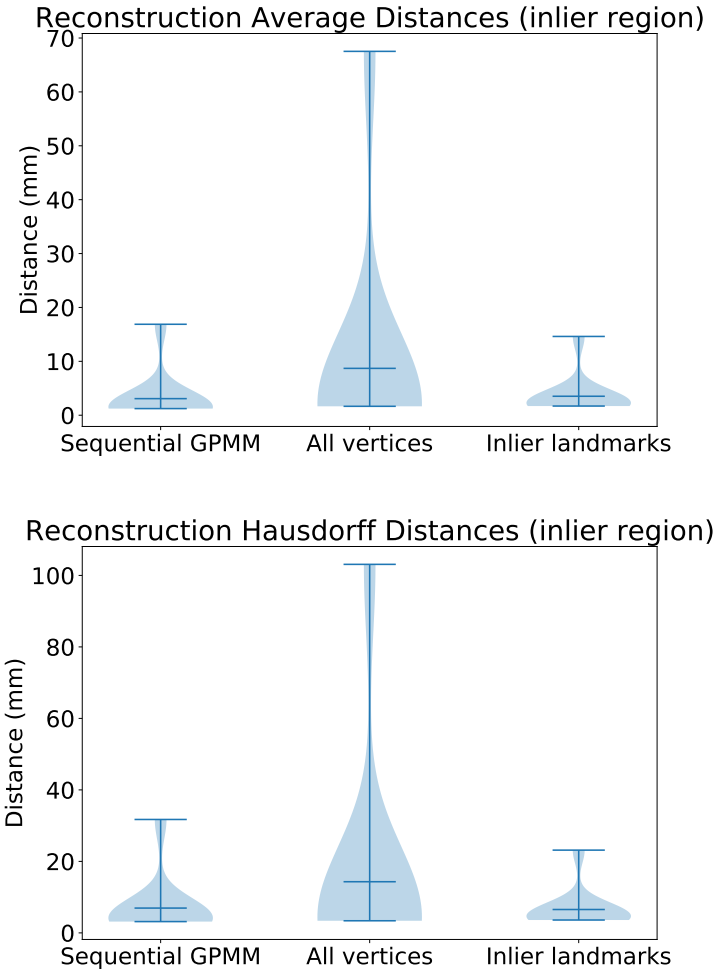


Figure 6.5: Violin plots comparing the average and Hausdorff distances of the 10 partial liver reconstructions using the proposed sequential GPMM and the traditional GPMM regression approahes. The distances are computed in the inlier region, which is the domain provided by the target.

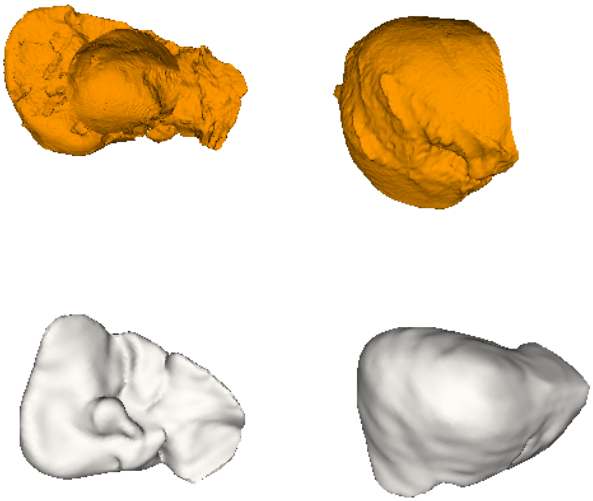


Figure 6.6: Top row: Two example partial livers from the Shape 2015 dataset: liver 135 (left), with a missing surface on the posterior side, and liver 132 (right), with a missing part from the left lobe. Bottom row: The reference mesh used to build the full liver SSM is shown for reference from the posterior (left) and anterior (right) view).

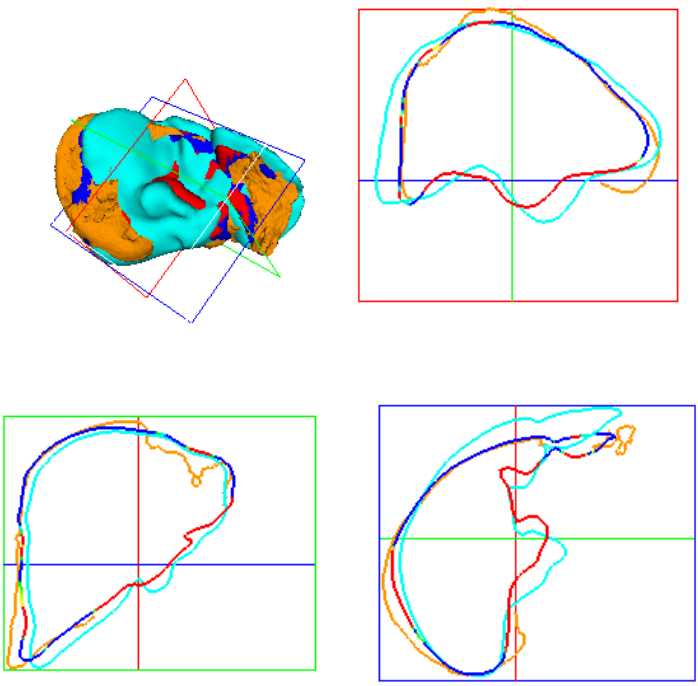


Figure 6.7: Sequential GPMM (blue and red) and GPMM regression (light blue) reconstructions of partial liver number 135 from the Shape 2015 dataset (orange). The sequential GPMM gives an average distance and Hausdorff distance of 1.84 mm and 14.44 mm, compared to 4.82 mm and 17.20 mm respectively.

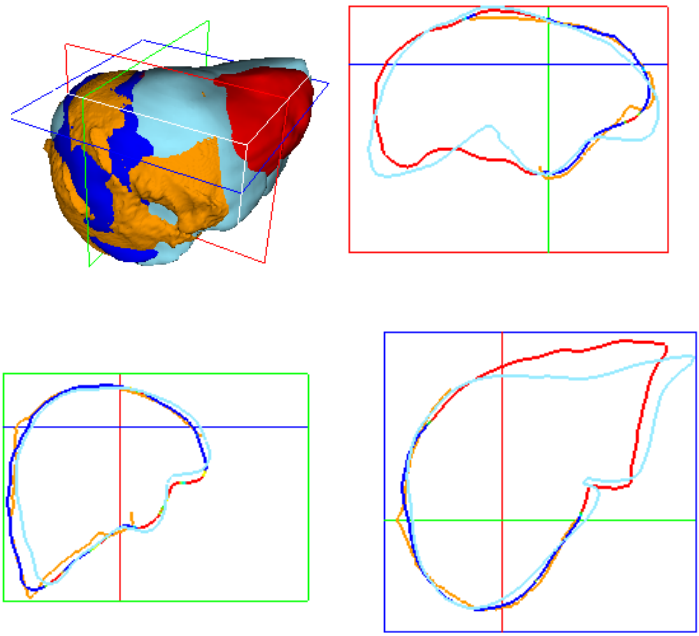


Figure 6.8: Sequential GPMM (blue and red) and GPMM regression (light blue) reconstructions of partial liver number 132 from the Shape 2015 dataset (orange). The sequential GPMM gives an average distance and Hausdorff distance of 1.76 mm and 15.80 mm, compared to 3.31 mm and 13.27 mm respectively.

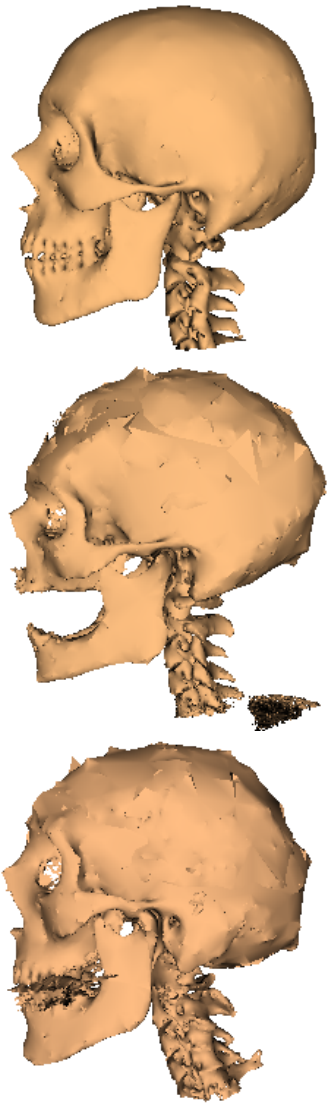


Figure 6.9: Examples from the forensics dataset: targets 3, 38 and 24.

The reconstruction results in the outlier regions are not calculated because the ground truth shapes are not available. The approaches Sequential GPMM and All vertices appear to perform similarly in the inlier reconstruction task, with Sequential GPMM only showing improvements in the mean values of the average distances. This is due to the evaluation approach, which computes the distances between the inlier region vertices and their closest points on the target without relying on ground truth correspondence. Some visualizations are provided next to better explain the advantages of the Sequential GPMM. Three examples are shown for target numbers 3, 24 and 38 shown in figure 6.9. From the full skull targets (orange), the mandible with teeth is reconstructed. The reconstructions with the predicted labels are shown from figure 6.11 to figure 6.16. The reconstructions obtained with the proposed sequential GPMM are colored with blue and red, where the colors indicate the vertex labels for inliers and outliers respectively. The reconstruction obtained using the full model reference for fitting is in light blue. As expected from a robust method, the sequential GPMM reconstruction (blue for inliers, red for outliers) of the target (orange) is more accurate than the traditional GP regression (light blue). Because it excludes outlier observations from inference, it is able the reconstruct the available inlier regions with higher accuracy.
A robust simultaneous reconstruction and detection algorithm can be helpful for cranio-maxillofacial surgery planning [23] and teeth detection and counting [74]. As proof-of-concept, the label map obtained from sequential GPMM is used to predict the teeth that are missing or whose surface information is lost due to artifacts on the target. This is automated by evaluating for each tooth the percentage of its vertices that are labeled as outliers, Percentages above 30% are labeled as missing. The tooth detections and the reconstruction distances are listed in the captions of the figures. Quantitative evaluation of this approach requires information about the missing teeth in each skull of the forensics dataset. Future work can include evaluation of the label map against expert segmentations. Evaluation of the individual tooth reconstruction quality can also be performed by evaluating the outlier region reconstruction distances, for example for post-mortem identification [75].

6.4.2 Femur reconstruction

The problems of the femur in figure 6.17 are that both ends are missing and that the distal and proximal ends of the femur are not attached due to the fracture. The first problem makes length estimation a challenge. Sequential

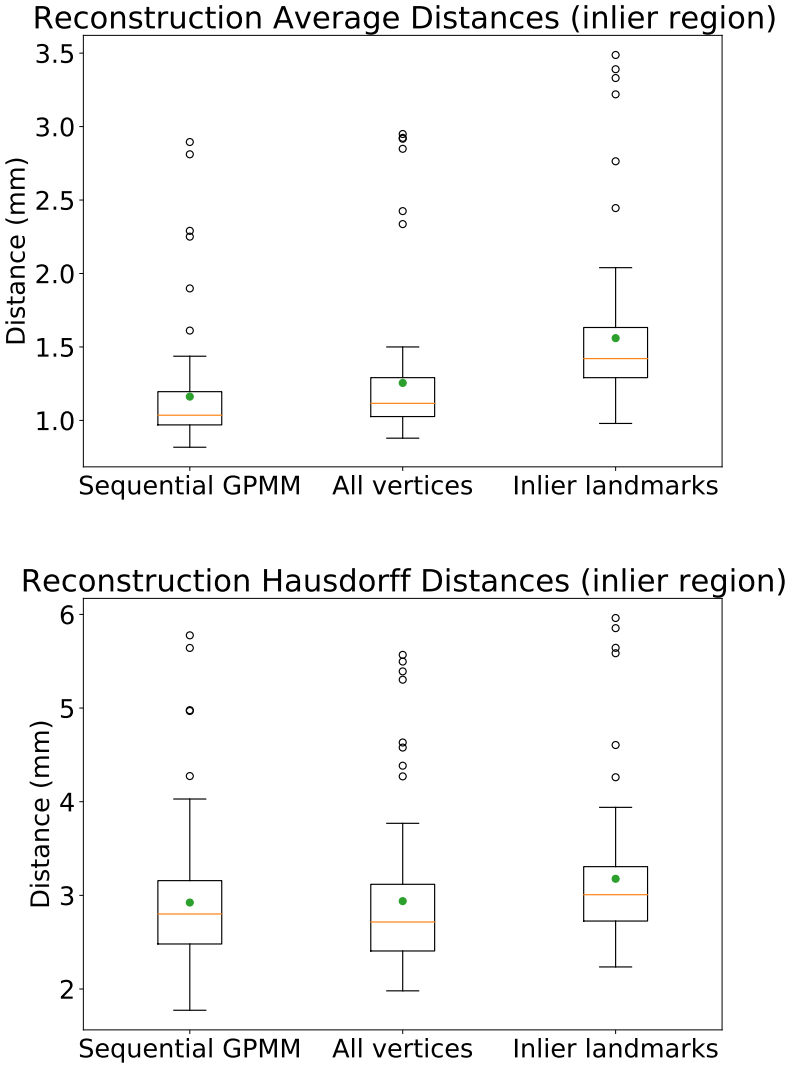


Figure 6.10: Boxplots summarizing the 86 mandible reconstruction results from the forensics dataset.

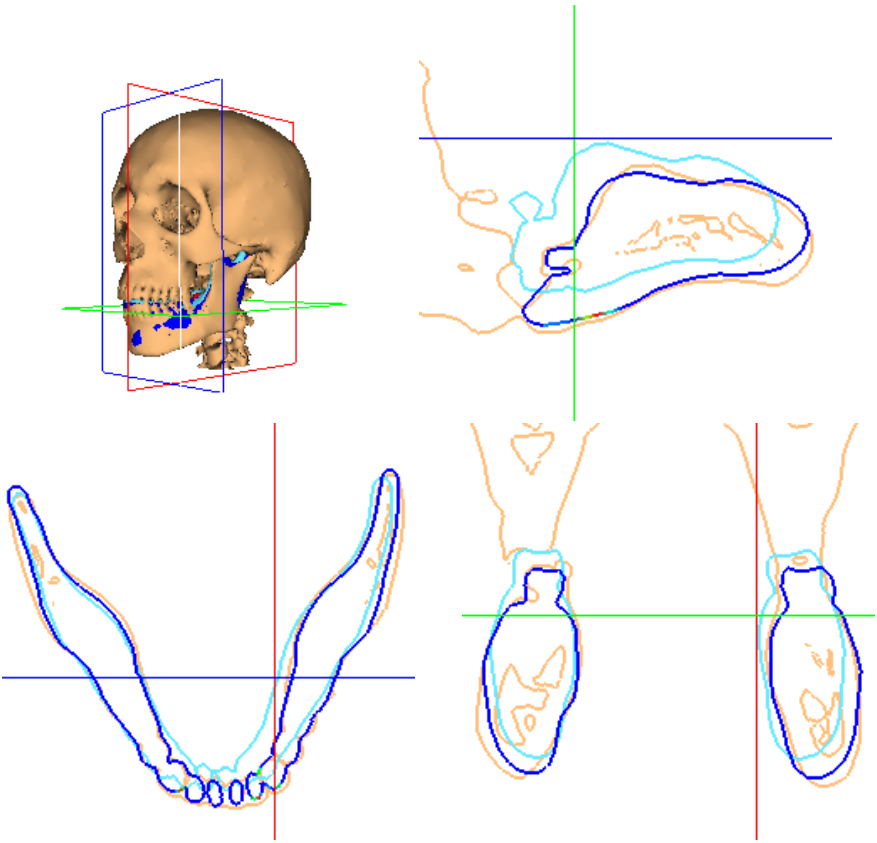


Figure 6.11: Example reconstruction from the forensics dataset on target number 3. The sequential GPMM reconstruction (blue for inliers, red for outliers) of the target (orange) is more accurate than the traditional GP regression (light blue), with average distance in the inlier region of 0.84 mm versus 1.28 mm. The regression quality in the inlier region shows improvement after outlier removal. The planes indicate the locations of the slices, taken through a healthy region on the target (top right: red plane, bottom left: green plane, bottom right: blue plane).

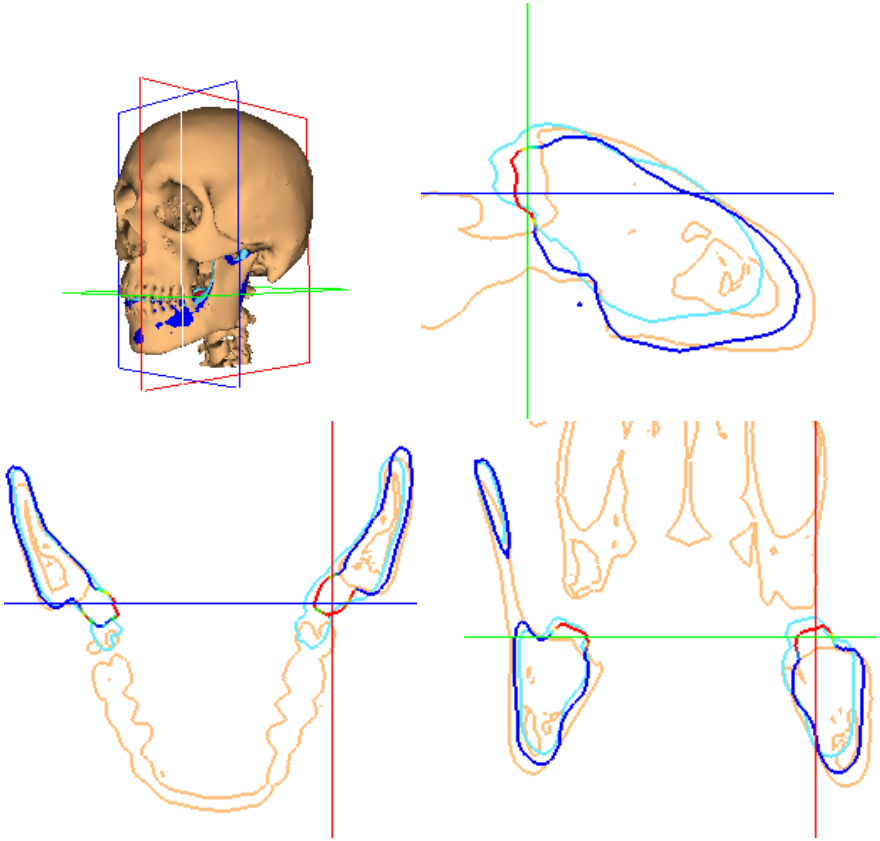


Figure 6.12: Example reconstruction from the forensics dataset on the same target as in image 6.11. Unlike the previous figure, the planes are taken through a pathological region on the target (top right: red plane, bottom left: green plane, bottom right: blue plane). Qualitative comparison of the missing tooth reconstruction shows that the proposed sequential GPMM (red tooth in the image slices) is able to more accurately estimate the location of the tooth than the traditional fitting method. The proposed teeth detection strategy proposes that the third molar (wisdom tooth) is missing on both left and right sides.

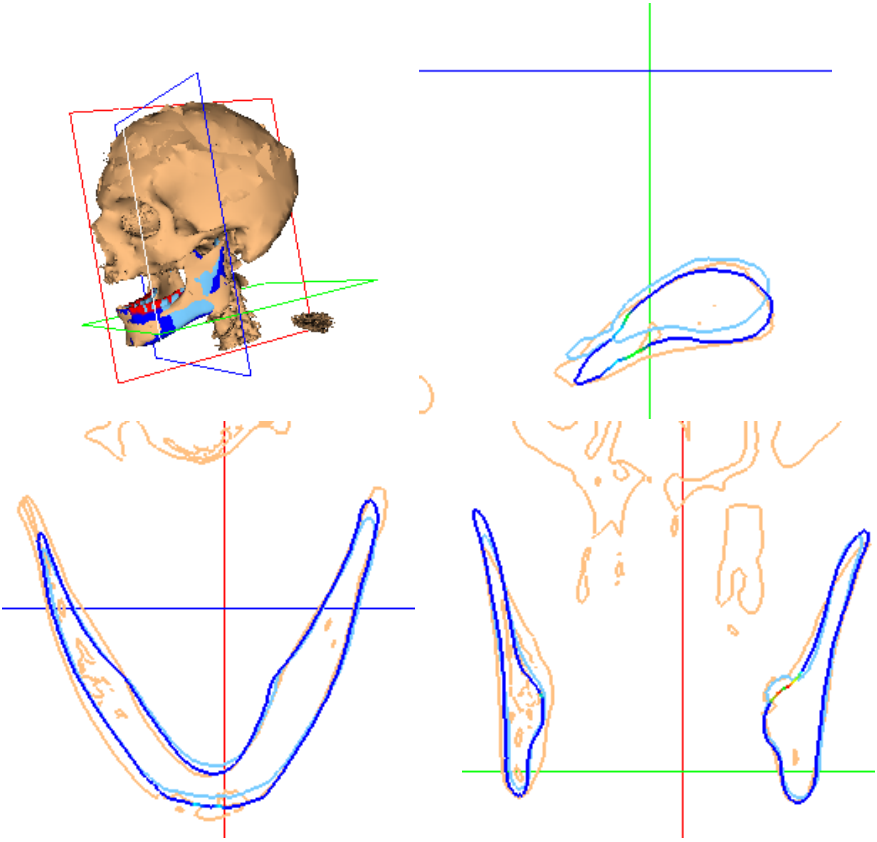


Figure 6.13: Example reconstruction from the forensics dataset on target number 38. The sequential GPMM reconstruction (blue for inliers, red for outliers) of the target (orange) is more accurate than the traditional GP regression (light blue), with average distance in the inlier region of 0.98 mm versus 1.04 mm. The planes indicate the locations of the slices, taken through a healthy region on the target (top right: red plane, bottom left: green plane, bottom right: blue plane).

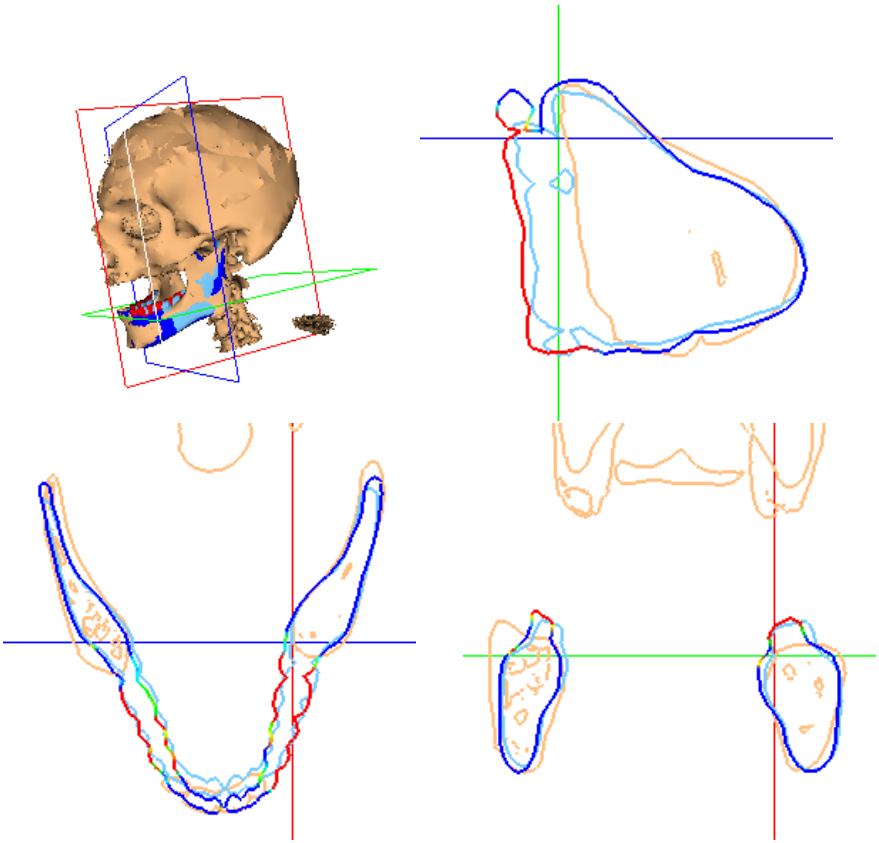


Figure 6.14: Example reconstruction from the forensics dataset on the same target as in image 6.13. Unlike the previous figure, the planes are taken through a pathological region on the target (top right: red plane, bottom left: green plane, bottom right: blue plane). Qualitative comparison shows that the proposed sequential GPMM (red tooth in the image slices) is able to more accurately estimate the location of the tooth than the traditional fitting method. The proposed teeth detection strategy proposes that 10 of the 16 mandible teeth are missing.

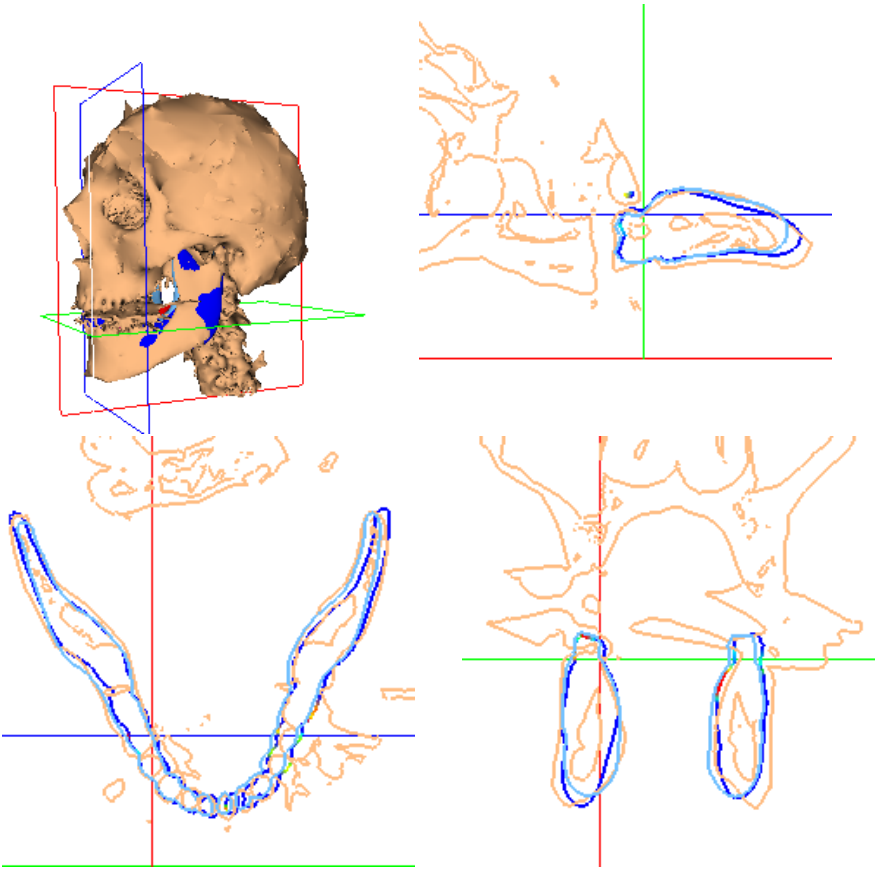


Figure 6.15: Example reconstruction from the forensics dataset on target number 24. The sequential GPMM reconstruction (blue for inliers, red for outliers) of the target (orange) is more accurate than the traditional GP regression (light blue), with average distance in the inlier region of 0.88 mm versus 1.05 mm. The planes indicate the locations of the slices, taken through a healthy region on the target (top right: red plane, bottom left: green plane, bottom right: blue plane).

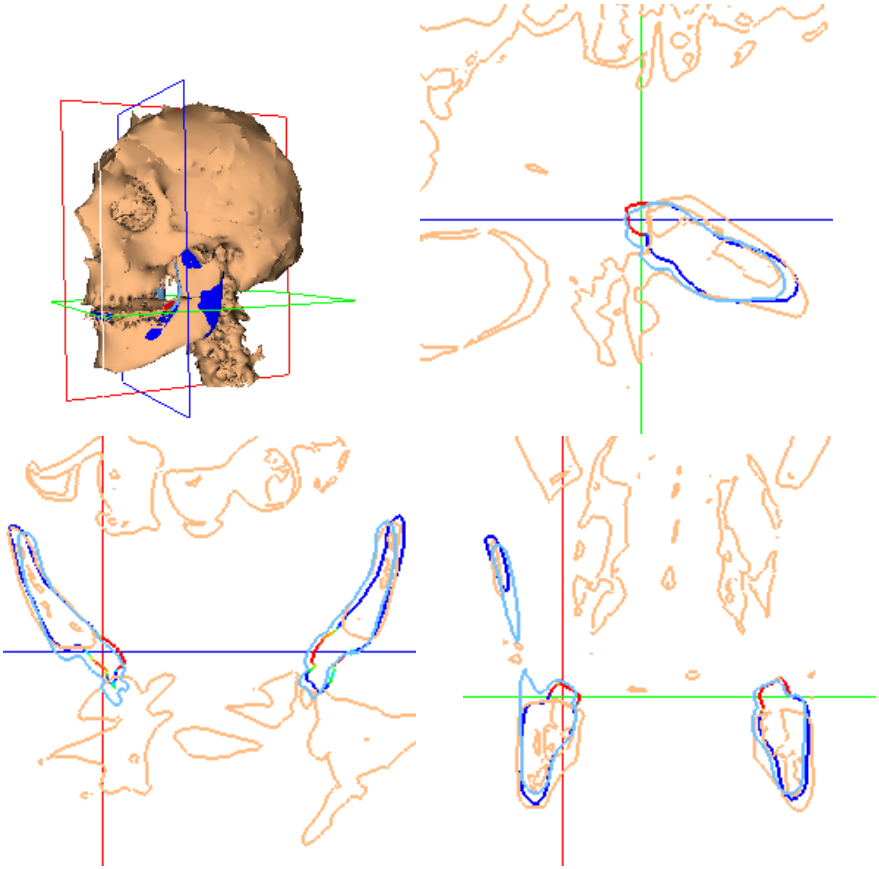


Figure 6.16: Example reconstruction from the forensics dataset on the same target as in image 6.15. Unlike the previous figure, the planes are taken through a pathological region on the target (top right: red plane, bottom left: green plane, bottom right: blue plane). The proposed teeth detection strategy labels the two wisdom teeth and the first molar on the right as missing.

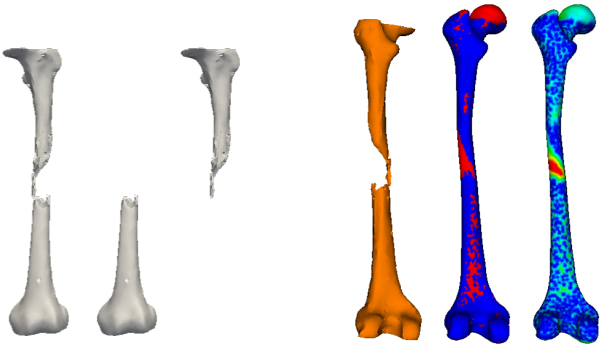


Figure 6.17: Fractured femur (white, left) with missing extremities as well as potential misalignment due to the missing attachment between the distal (white, center) and proximal (white, right) ends. The anterior side of the target is shown (orange) alongside the final pathology label map on the reconstruction and the Mahalanobis distance map used for obtaining the threshold and binary labels.

GPMM ensures the shape model does not shrink, because the parts on the reference that are missing on the target do not get falsely matched to the target extremities. The second problem might cause axial misalignment, which can corrupt the reconstruction.
To assess these problems, different reconstructions are performed. The first one performs reconstruction on the full target that has been extracted from the CT image. The alignment and sequential GPMM initialization are based on 6 landmarks, clicked on the distal and proximal ends. The second one performs reconstruction on the distal end of the target. The distal end of the target is easily obtained by separating the mesh based on connectivity, resulting in the proximal and distal ends visualized in figure 6.17. From the full target landmarks, only the 4 on the distal end are chosen for alignment and initialization of the region-growing algorithm. The reconstructions are then compared to evaluate whether there is misalignment in the target. The overlay comparison is seen in figure 6.18. The misalignment can be deduced after comparing the reconstructions of the femoral head, highlighted in the proximal end image slice (blue plane in the top row).
This leaves the question of whether sequential GPMM can be used to suggest the presence of a misalignment. This is tested on this clinical example. Both proximal and distal ends are provided as the target. Initialization is

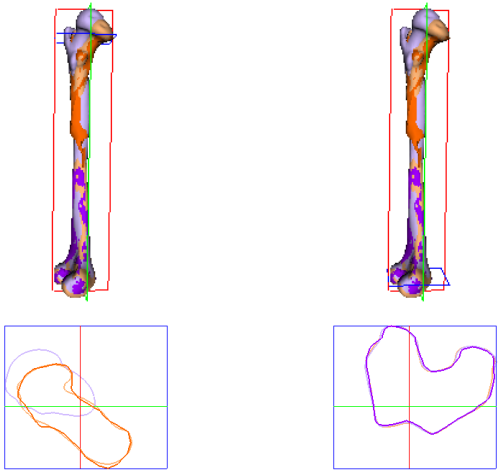


Figure 6.18: Fitting the femur model to the full target (dark orange) or the distal part of the target (dark purple) gives two outputs, in light orange and purple respectively. The first column shows a slice through the proximal end, while the second a slice through the distal end, highlighting the reconstruction differences in the proximal end.

performed with landmarks on the distal end only. The label map of the pathology estimated by sequential GPMM is shown alongside the target in figure 6.17. The pathology labels on the shaft of the distal part can be an indicator of the misalignment that is present in the target. The Mahalanobis distances used to obtain the threshold and label map is also shown in the figure. The distances vary along the shaft and do not follow a smooth increase as seen in the missing part in the center of the shaft. The proximal and distal parts are healthy and can be reconstructed and correctly labeled if fitted separately. However, taken together, the reconstruction is deformed and the label map of each end includes false outliers. The label map quality can be used to indicate reconstruction uncertainty and hint at potential pathologies that have not been considered by the proposed algorithm. To apply this in a clinical setting, however, a thorough investigation should include expert quantification of the axial misalignment, comparison of the reconstruction to the ground truth contralateral femur, and evaluation of the different pathology regions predicted by the method based on the Mahalanobis distances. A task-specific correspondence function could be considered in the testing phase, to increase robustness by eliminating observations that cause deformations towards misaligned regions.

Chapter 7

Future Work

In addition to the clinical applications mentioned in section 6.4, the following ideas can be used to guide further developments of the sequential GPMM algorithm. The first two are task-specific changes that can be included to facilitate initialization or improve reconstruction results of the pathology region. The third suggestion aims to replace the used GPMM to extend the proposed method to the image domain.

7.1 Landmark Initialization

Landmarks used in the current implementations of the proposed algorithms are provided by the user, which is possible because only a few are needed to obtain a rough model-target alignment and to act as seeds. Nevertheless, landmarks can be automatically generated for the model reference and novel targets through different approaches. A previous master thesis from the group has already implemented the heat kernel signature [76] method for meshes in Scalismo. More recent deep learning based approaches can be used to provide an initial set of landmarks [50]. The performance of both strategies would first need to be evaluated on pathological data, and prediction uncertainty can be integrated into the GP regression pipeline through the Gaussian noise assumed on each landmark observation.

7.2 Sample Selection

The thesis showed two different approaches that can be used to obtain the posterior distribution describing the healthy distribution of a pathological target: an approximate solution in chapter 3, and a closed-form one in chapter 4. The MAP solution is then chosen as the shape reconstruction, and the label map is projected onto it. The results on clinical data in chapter 6 reveal that this solution is not always sufficient for the outlier region prediction. For example, the prediction shown on the second example from the Autoimplant dataset in figure 6.2 does not reconstruct the implant accurately. Instead of taking the MAP solution, which was the mean of the PPD obtained at the end of the GPMM regression, it is possible to choose another sample from the PPD that minimizes a case-specific expected loss. Section 2.4 from the book [5] describes how an expected loss can be minimized when choosing a reconstruction from the PPD using decision theory for regression. When the loss function is symmetric and the PPD is Gaussian, the chosen solution is always the mean of the PPD, as was the case with our MAP solution. Choosing an asymmetric loss function to minimize would enable more flexibility with choosing the best reconstruction of the outlier region. The loss function can be shape family specific or target specific. Alternatively, case-specific corrections can be introduced to account for thickness inaccuracies [77], diameter mismatch between the implant prediction and the inlier region using a step-off loss [78], or other reconstruction requirements defined by relevant medical experts.

7.3 Model Extensions

The outlier detection solutions for pathology detection have only been implemented for shape deformations. However, they are general solutions and can be extended to include other pathology types. For intensity-based pathologies or inner-volume pathologies, the sequential GPMM algorithm can be implemented with a GPMM built on the image domain. The image registration process would be based on the formulation presented in [3]. The extension would enable direct comparison to methods that perform detection in the image domain [27, 28, 79] on datasets such as the KITS challenge previously discussed in chapter 6 or the BraTS dataset [80–82]. Otherwise, a shape model based on tetrahedral meshes with an inner grid [83] or using a statistical shape and intensity model (SSIM) formulation instead of the standard surface-mesh GPMM [84] can be used in order to store information about the inner-volume of the shape. Inner-volume vertices augmented with

intensity information or 3D image-based models can enable inner-volume shape and intensity pathology detection.
The noise model can also be extended to include correlated noise, which is currently being developed in another thesis in the group. The current likelihood model used for approximating the predictive posterior distribution in the forward SSM approach relies on the assumption of independent Gaussian noise on the observations. Including the correlated noise model into the proposed sequential algorithm should enable more accurate F1 scores, by decreasing the number of false labels that are in fact due to correlated noise.

Chapter 8

Conclusion

The thesis presented shape fitting to targets with pathologies as an outlier detection problem. Outlier detection provides a consistent approach to dealing with outliers. Outliers are defined as observations that are not samples from the underlying model with its noise assumption. The problem becomes a one-class classification problem. The goal is to obtain a distribution for the inliers inliers, which upon thresholding can be used to identify outliers. The thesis showed that is possible to use outlier detection for pathology detection in shape modeling. Pathologies are defined in terms of outlier detection notation in chapter 1, after which the definition is used to show how to describe previous pathology detection algorithms developed for shape modeling in terms of outlier detection. Two different solutions are proposed to obtain the predictive posterior distribution for targets with pathologies. The first one in chapter 3 provides an approximate solution through sampling, while the second one in chapter 4 provides the closed-form solution. By combining region-growing with outlier detection, the closed-form predictive posterior distribution was obtained starting from a GPMM conditioned on inlier landmarks. Chapter 5 compared the performance of the proposed outlier detection solutions to previous shape reconstruction methods. This was followed by an evaluation of the fitting sensitivity to hyperparameters as well as a breakdown point analysis. The thesis covered shape pathologies which can be described in terms of deformation vectors that deviate significantly from their expected distribution under the GPMM. The main future work direction is extending the model to intensity pathologies, where pathologies would be described in terms of deviations from the expected intensity distributions defined by the GPMM.

This would enable detection of pathologies in volumetric images that do not cause surface deformations as well as inner-volume pathologies. The thesis reveals the possibility of using general pathology definitions in terms of outlier detection in shape modeling applications. This comes with the benefit of applying such general solutions to different shape families without prior information or definitions of pathologies. This was shown in chapter 6 on example targets from the: KITS kidney tumor detection challenge, AutoImplant cranium implant reconstruction challenge, Shape liver reconstruction challenge, in-house femur and mandible examples. The success of the outlier detection definition of pathologies is an encouragement for the community, since it shows that pathology-specific training sets are not necessary. Instead of relying on data-driven segmentation and detection algorithms trained on pathological examples, an attainable training set that describes healthy data can be used. Future work should focus on pathology-agnostic approaches and aim towards general notation, datasets and metrics that would enable the comparison and clinical evaluation of such algorithms.

Appendix A

Chapter publications

- Chapters 3 and 5 are based on the publication at the LABELS workshop from MICCAI 2019 [45]. The authenticated publication is available online at https://doi.org/10.1007/978-3-030-33642-4_2.
- Chapters 4, 5, and 6 are based on the publication at MICCAI 2021 [66]. The authenticated publication is available online at https://doi.org/10.1007/978-3-030-87240-3_41.
- The comparison to other robust Gaussian process regression approaches in chapter 4 and appendix B was presented at ICORS 2021.

Appendix B

Outlier cases for Robust Gaussian Process Inference

This appendix shows results for the experiments discussed in section 4.4. Compared to the closed-form methods, the approximate methods (bowoGP, tlikeGP) are less accurate in the reconstruction of the inlier regions of the function and require longer runtimes. The proposed method (seqGP) shows reconstruction results that are as good as the robust closed-form solutions (hubGP, itGP), with the additional advantange of reduced runtime and robust reconstruction under smooth outlier cases. The outlier cases presented here are:

- No outliers
- Rare outliers (5% outliers, outlier noise $\mathcal{N}(0, 1)$)
- Fiducial outliers (15% outliers, outlier noise $\mathcal{N}(0, 1)$)
- Abundant outliers (45% outliers, outlier noise $\mathcal{N}(0, 1)$)
- Skewed outliers (15% outliers, outlier noise $\mathcal{N}(2, 1)$)
- Uniform outliers (30% outliers, outlier noise $\mathcal{U}(-3, 3)$)
- Student distribution with 3 degrees of freedom outliers (100% outliers, outlier noise t_3 distributed)
- Smooth missing data outliers (hole: 10% outliers, outlier observations are removed from the provided training set)

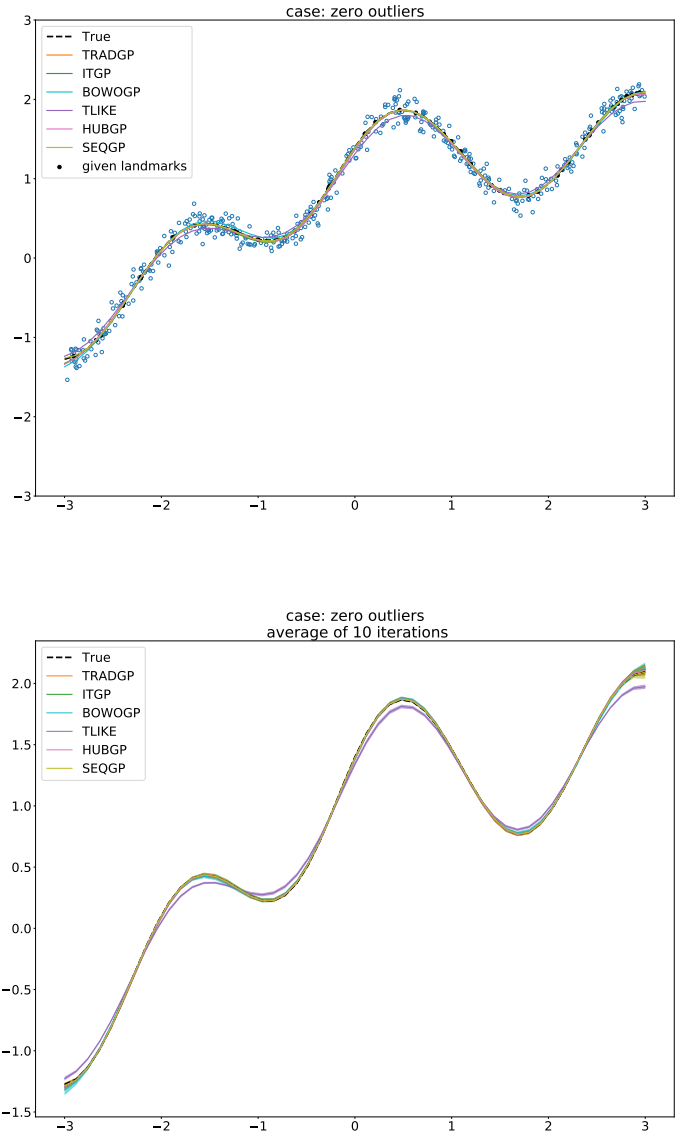


Figure B.1: Case: No outliers. Top: mean of the inferred GP obtained from the different robust GP inference methods. Bottom: Average mean and standard deviation of the inferred GP mean from 10 different iterations.

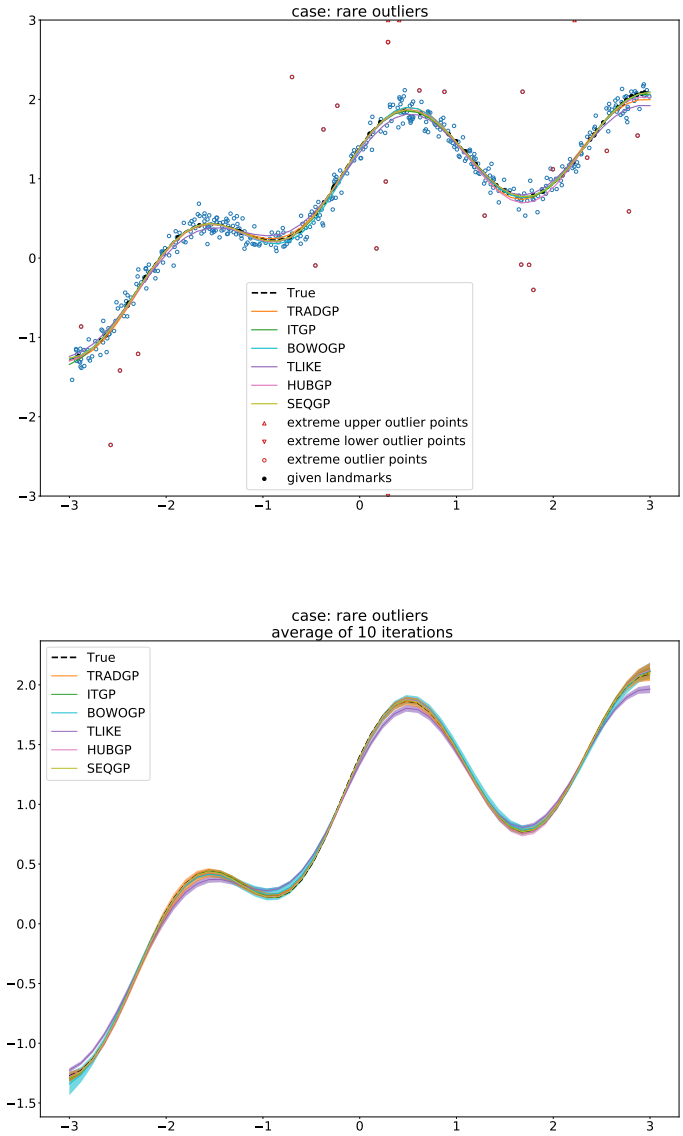


Figure B.2: Case: Rare outliers. Top: mean of the inferred GP obtained from the different robust GP inference methods. Bottom: Average mean and standard deviation of the inferred GP mean from 10 different iterations.

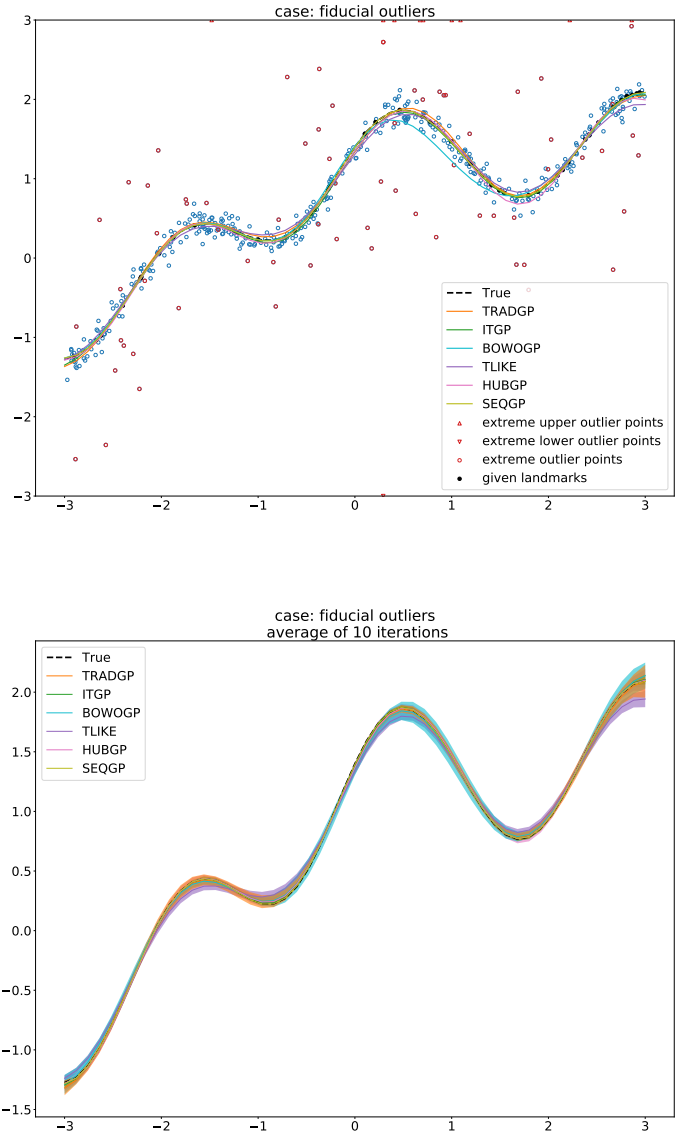


Figure B.3: Case: Fiducial outliers. Top: mean of the inferred GP obtained from the different robust GP inference methods. Bottom: Average mean and standard deviation of the inferred GP mean from 10 different iterations.

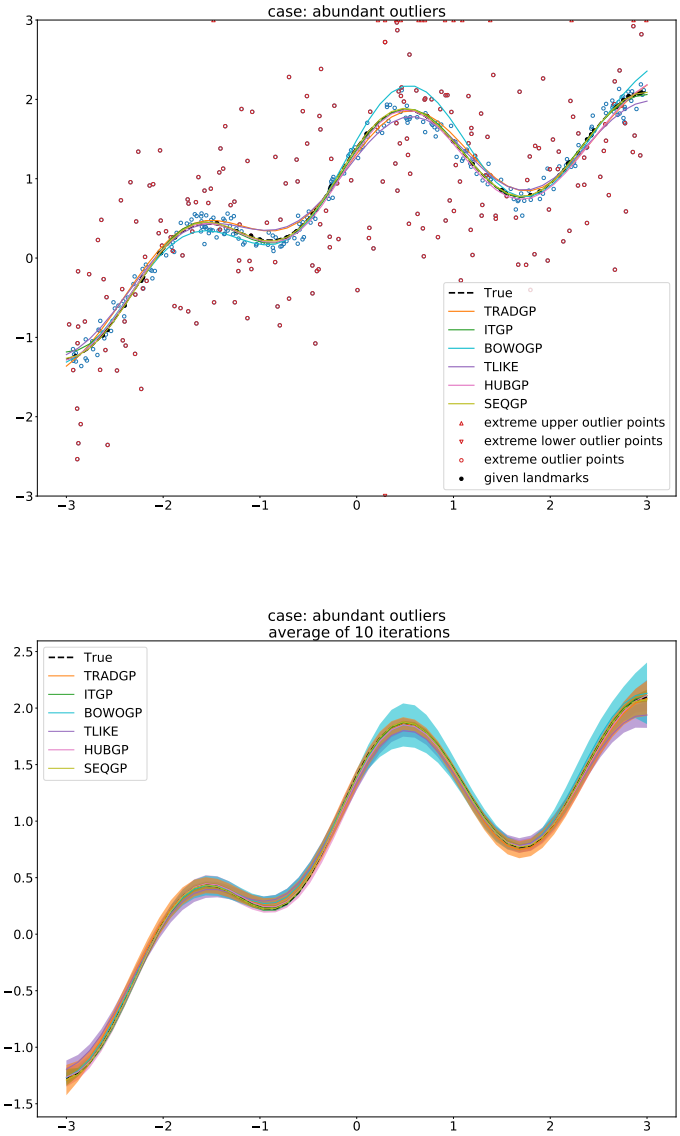


Figure B.4: Case: Abundant outliers. Top: mean of the inferred GP obtained from the different robust GP inference methods. Bottom: Average mean and standard deviation of the inferred GP mean from 10 different iterations..

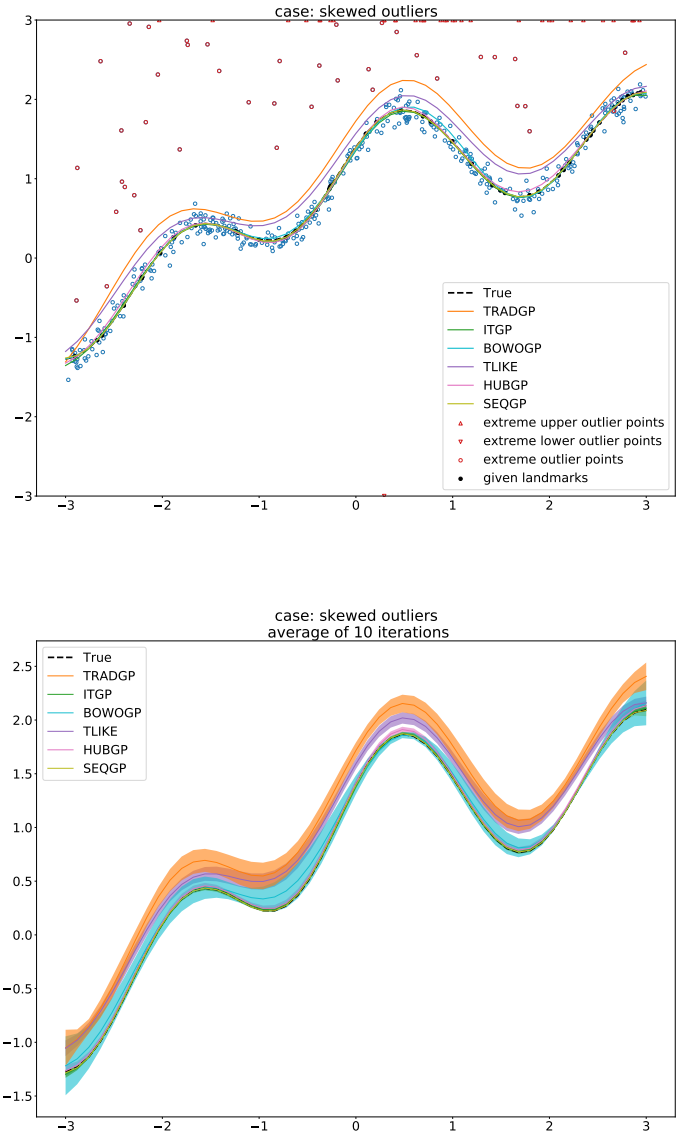


Figure B.5: Case: Skewed outliers. Top: mean of the inferred GP obtained from the different robust GP inference methods. Bottom: Average mean and standard deviation of the inferred GP mean from 10 different iterations.

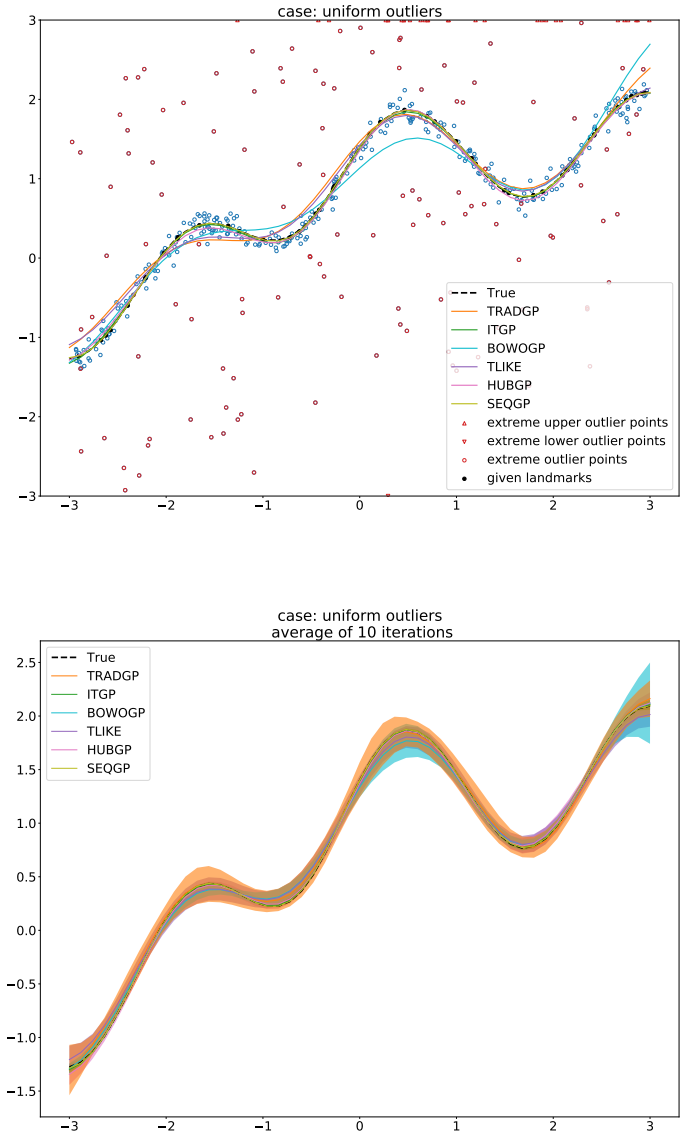


Figure B.6: Case: Uniform outliers. Top: mean of the inferred GP obtained from the different robust GP inference methods. Bottom: Average mean and standard deviation of the inferred GP mean from 10 different iterations.

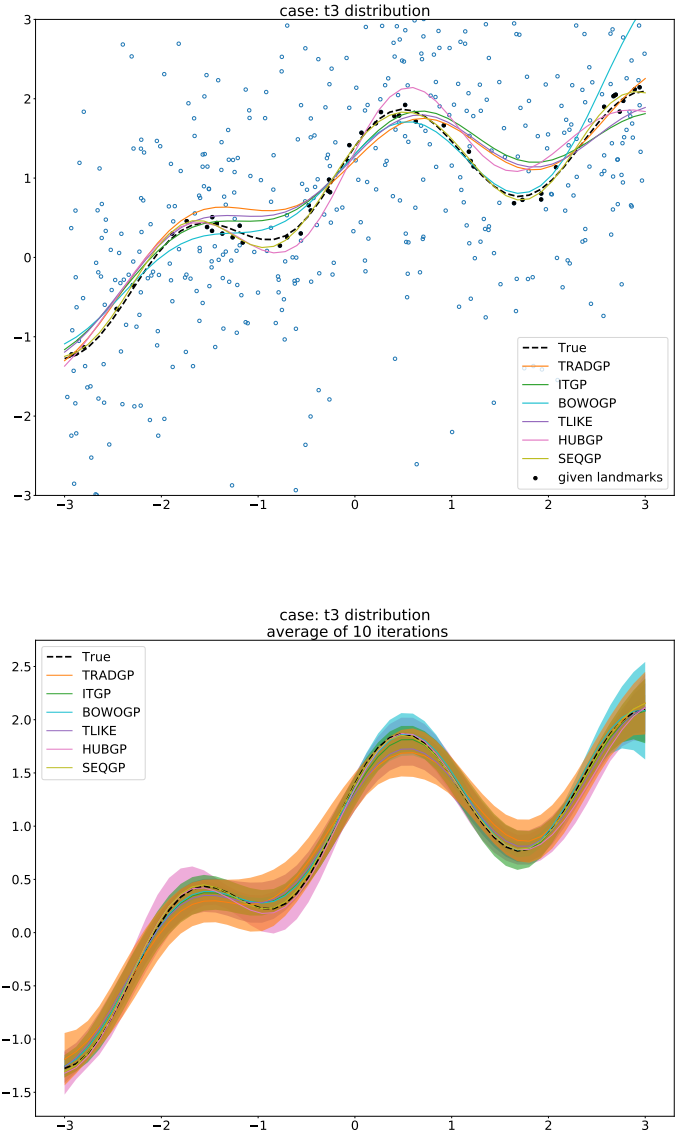


Figure B.7: Case: Student-distributed with 3 degrees of freedom noise. Top: mean of the inferred GP obtained from the different robust GP inference methods. Bottom: Average mean and standard deviation of the inferred GP mean from 10 different iterations.

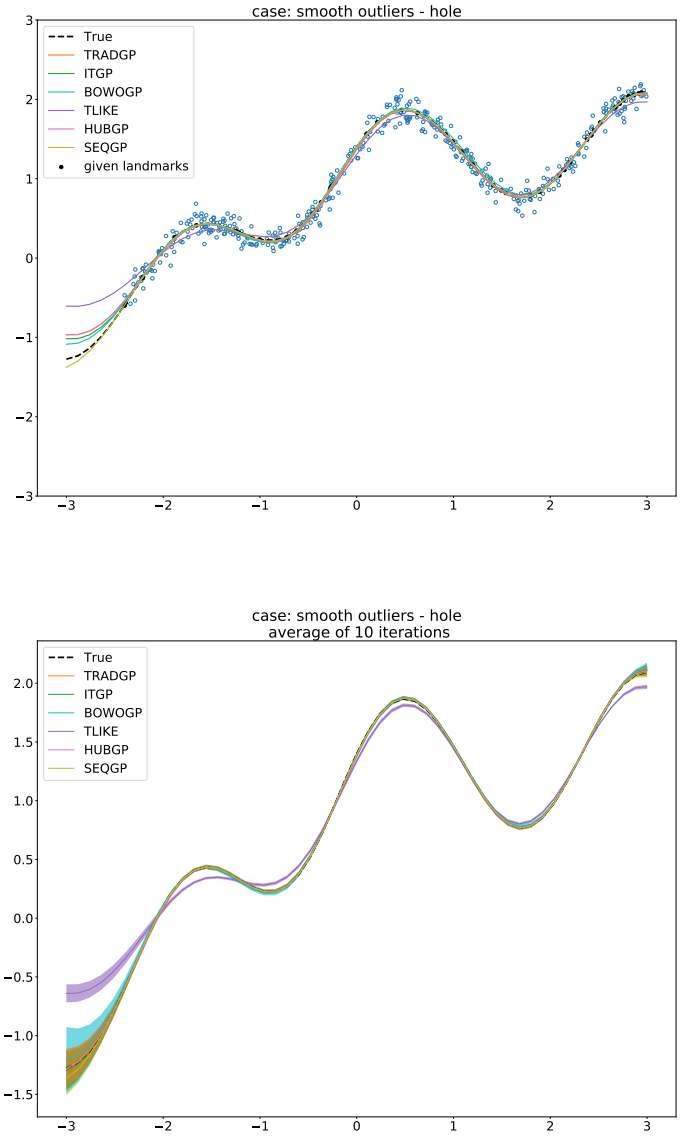


Figure B.8: Case: Smooth hole outliers. Top: mean of the inferred GP obtained from the different robust GP inference methods. Bottom: Average mean and standard deviation of the inferred GP mean from 10 different iterations.

Appendix C

Analysis for Forward SSM

C.0.1 Model rank and number of landmarks

The model rank does not influence the results, which is revealed by the similar results obtained for the different metrics when the model rank value changes in figure C.1. As for the number of landmarks used for initialization, the results are plotted for when the number of landmarks is $10, 5, 3$ and 1. Comparing the 10 landmark case in figure C.1 to the other cases in figures C.2, C.3, and C.4, reveals that a lower initial number of landmarks improves the F1 score, especially as the size of the outlier region increases. This is due to the increased flexibility remaining in the posterior model conditioned on the landmarks and used to initiate the sampling process.

C.0.2 Mesh density

Similar to the model rank, the reference mesh density does not influence the detection and reconstruction results, shown in figure C.5.

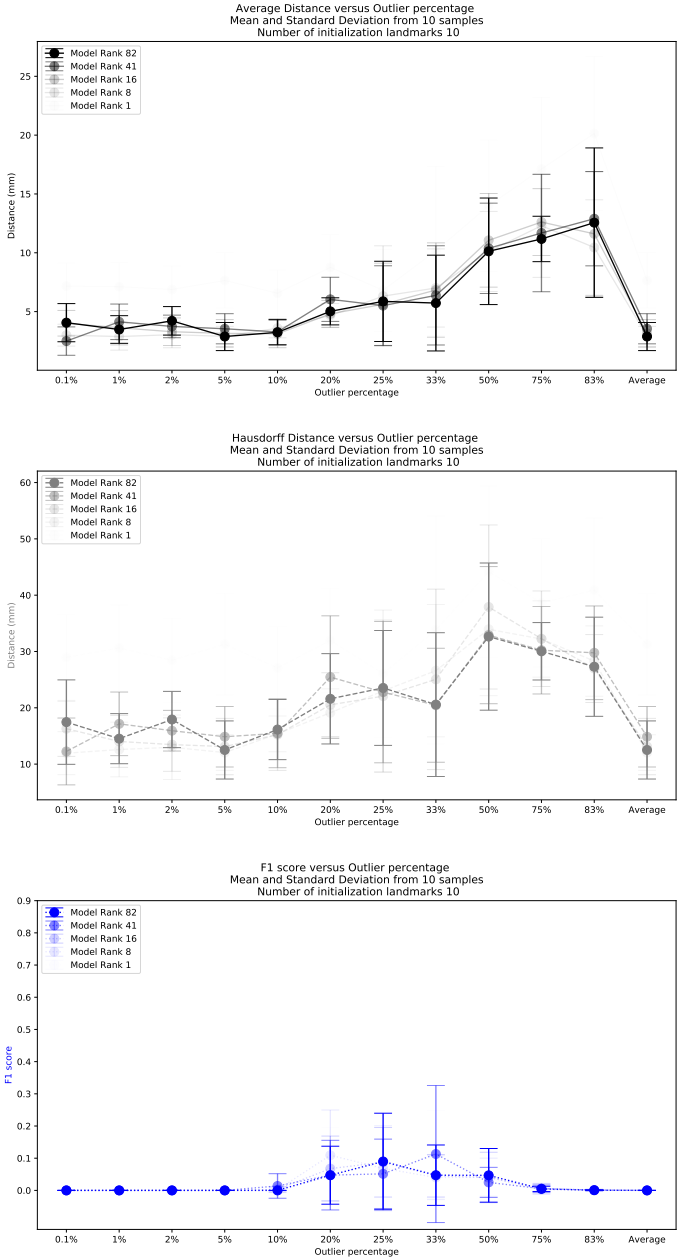


Figure C.1: Influence of model rank the on forwardSSM results. The color indicates the model rank. Reconstruction and detection metrics are plotted as the outlier percentage increases and are similar across ranks.

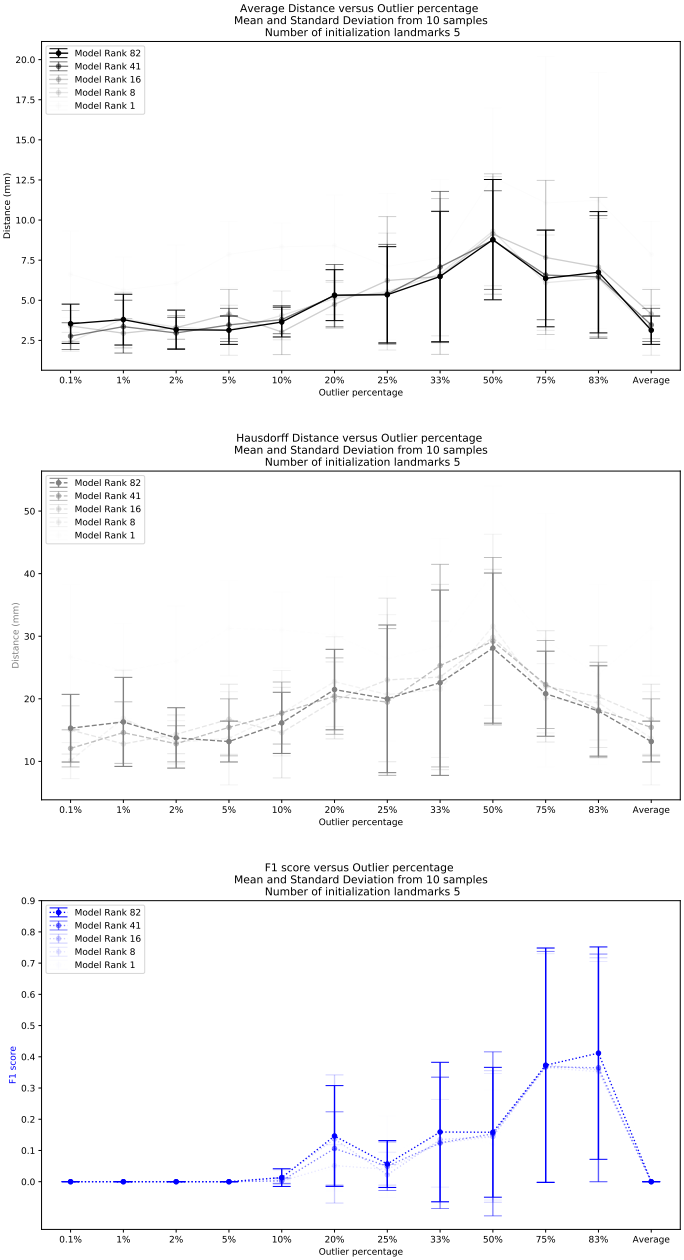


Figure C.2: Influence of model rank the on forwardSSM results. The color indicates the model rank. Reconstruction and detection metrics are plotted as the outlier percentage increases and are similar across ranks.

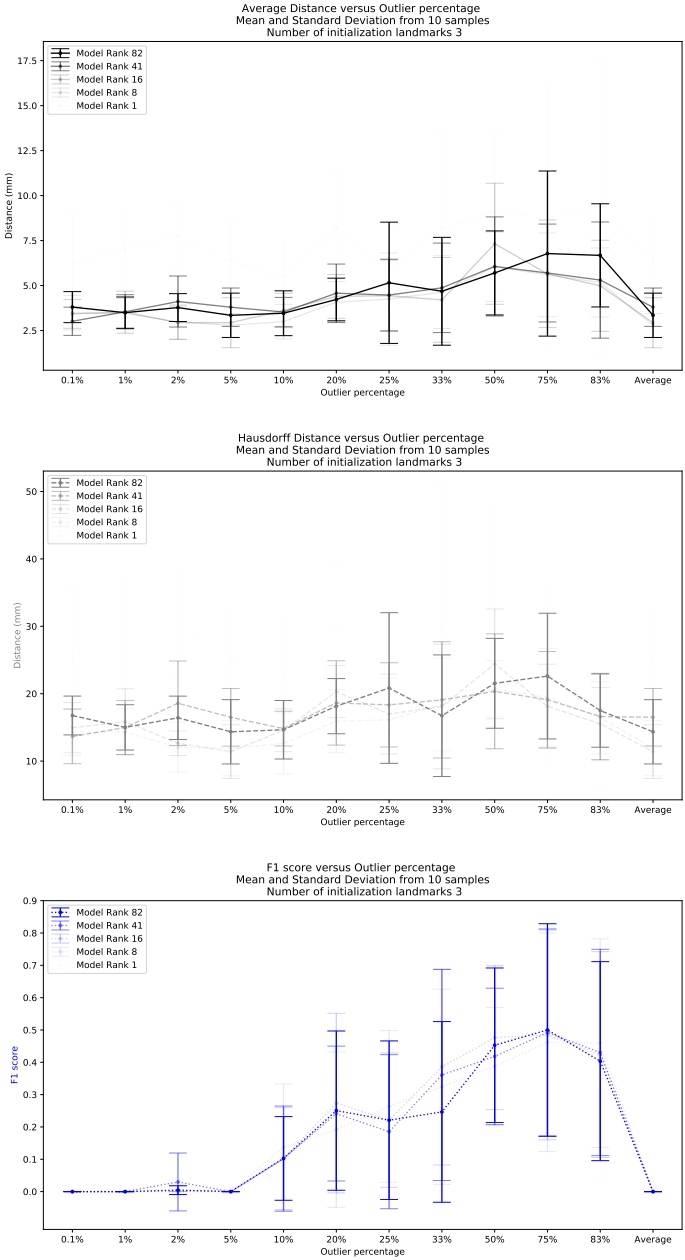


Figure C.3: Influence of model rank the on forwardSSM results. The color indicates the model rank. Reconstruction and detection metrics are plotted as the outlier percentage increases and are similar across ranks.

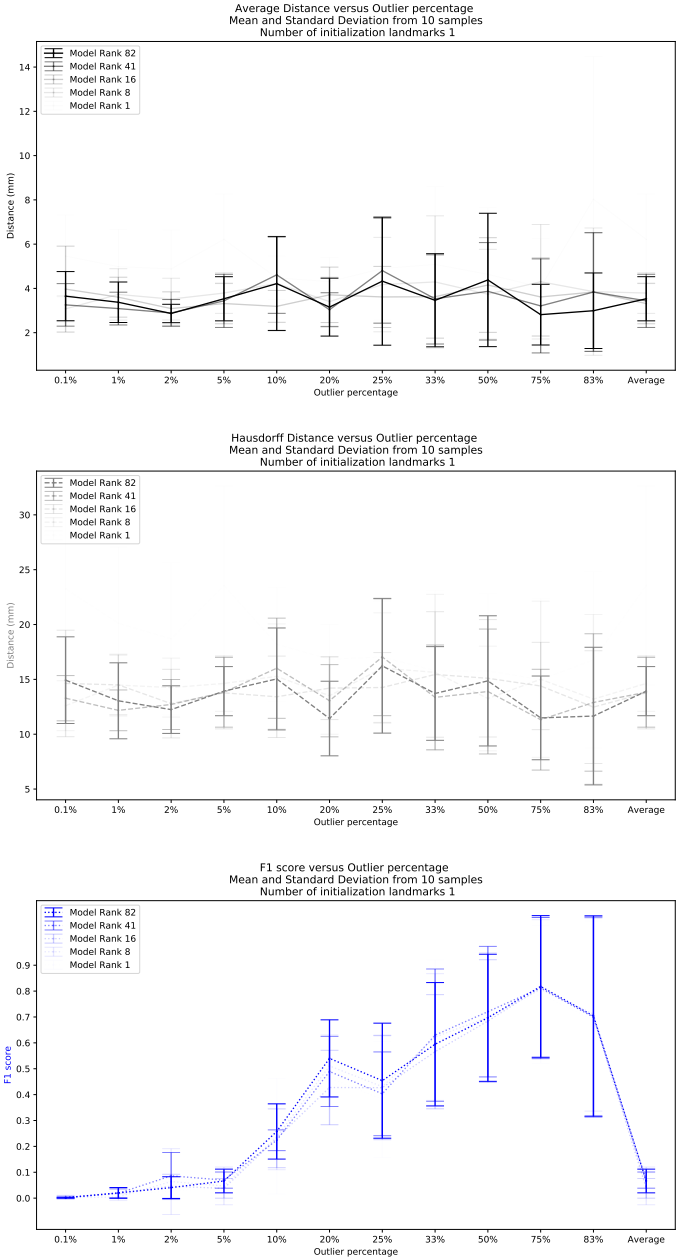


Figure C.4: Influence of model rank the on forwardSSM results. The color indicates the model rank. Reconstruction and detection metrics are plotted as the outlier percentage increases and are similar across ranks.

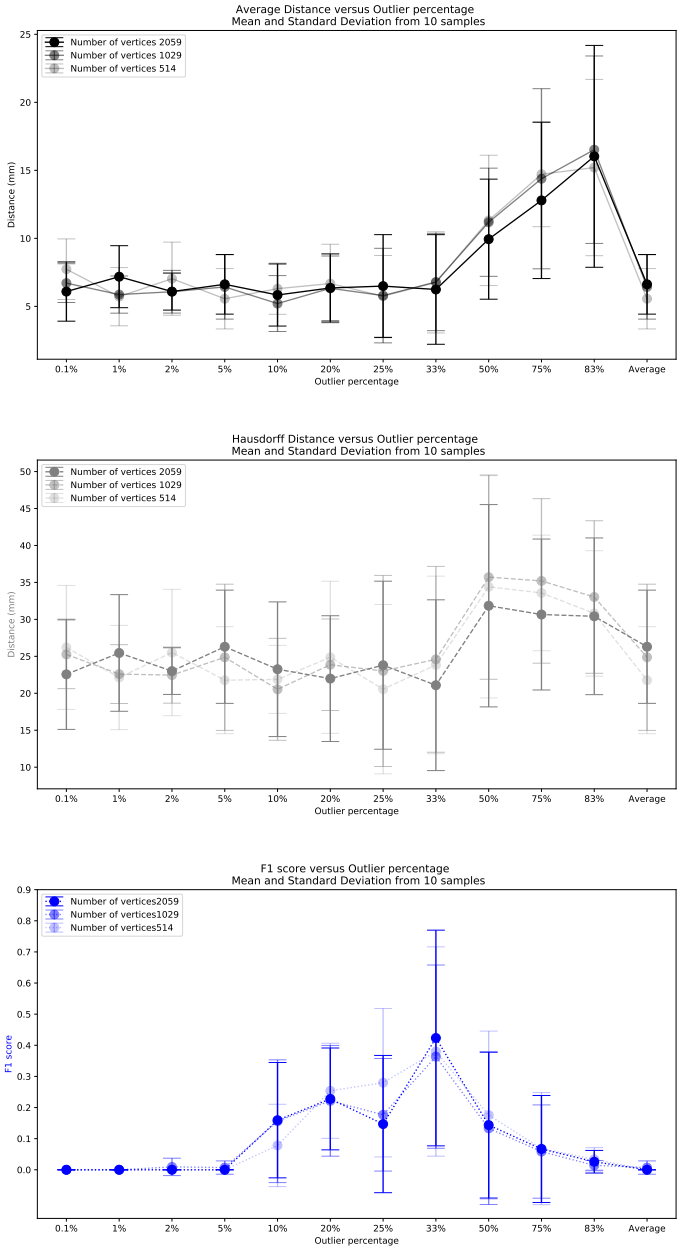


Figure C.5: Influence of mesh density on forwardSSM. The color indicates the reference mesh density. Reconstruction and detection metrics are plotted as the outlier percentage increases and are similar across densities.

Bibliography

[1] Ruben Martinez-Cantin, Kevin Tee, and Michael McCourt. Practical bayesian optimization in the presence of outliers. In *International Conference on Artificial Intelligence and Statistics*, pages 1722–1731. PMLR, 2018.

[2] Thomas Gerig, Andreas Morel-Forster, Clemens Blumer, Bernhard Egger, Marcel Luthi, Sandro Schönborn, and Thomas Vetter. Morphable face models-an open framework. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 75–82. IEEE, 2018.

[3] Marcel Lüthi, Thomas Gerig, Christoph Jud, and Thomas Vetter. Gaussian process morphable models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(8):1860–1873, Aug 2018.

[4] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[5] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian processes for machine learning*. Adaptive computation and machine learning. MIT Press, Cambridge, Mass, 2006. OCLC: ocm61285753.

[6] Zhao-Zhou Li, Lu Li, and Zhengyi Shao. Robust gaussian process regression based on iterative trimming. *arXiv preprint arXiv:2011.11057*, 2020.

[7] William H Press, Saul A Teukolsky, William T Vetterling, and Brian P Flannery. *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press, 2007.

[8] Andreas Ruckstuhl. Robust fitting of parametric models based on m-estimation. *Lecture notes*, page 40, 2014.

[9] Chiwoo Park, David J Borth, Nicholas S Wilson, Chad N Hunter, and Fritz J Friedersdorf. Robust gaussian process regression with a bias model. *arXiv preprint arXiv:2001.04639*, 2020.

[10] Pasi Jylänki, Jarno Vanhatalo, and Aki Vehtari. Robust gaussian process regression with a student-t likelihood. *Journal of Machine Learning Research*, 12(11), 2011.

[11] Brian McWilliams, Gabriel Krummenacher, Mario Lucic, and Joachim Buhmann. Fast and robust least squares estimation in corrupted linear models. In *Neural Information Processing Systems (NIPS)*, December 2014.

[12] Holger Trittenbach, Adrian Englhardt, and Klemens Böhm. An overview and a benchmark of active learning for outlier detection with one-class classifiers. *Expert Systems with Applications*, page 114372, 2020.

[13] Bernhard Egger, Sandro Schönborn, Andreas Schneider, Adam Kortylewski, Andreas Morel-Forster, Clemens Blumer, and Thomas Vetter. Occlusion-aware 3d morphable models and an illumination prior for face image analysis. *International Journal of Computer Vision*, 126(12):1269–1287, 2018.

[14] Thomas Gerig. *Gaussian Process Morphable Models for Spatially-Varying Multi-Scale Registration*. PhD thesis, University_of_Basel, 2021.

[15] Michael Kemmler, Erik Rodner, Esther-Sabrina Wacker, and Joachim Denzler. One-class classification with gaussian processes. *Pattern recognition*, 46(12):3507–3518, 2013.

[16] Mark Smith, Steven Reece, Stephen Roberts, Ioannis Psorakis, and Iead Rezek. Maritime abnormality detection using gaussian processes. *Knowledge and information systems*, 38(3):717–741, 2014.

[17] Eva Schnider, Antal Horváth, Georg Rauter, Azhar Zam, Magdalena Müller-Gerbl, and Philippe C Cattin. 3d segmentation networks for excessive numbers of classes: Distinct bone segmentation in upper bodies. In *International Workshop on Machine Learning in Medical Imaging*, pages 40–49. Springer, 2020.

[18] Bao H Do, Curtis Langlotz, and Christopher F Beaulieu. Bone tumor diagnosis using a naïve bayesian model of demographic and radiographic features. *Journal of digital imaging*, 30(5):640–647, 2017.

[19] David Carl Dahlin and K Krishnan Unni. Bone tumors: General aspects and data on 8,547 cases. 1986.

[20] Marco AF Pimentel, David A Clifton, Lei Clifton, and Lionel Tarassenko. A review of novelty detection. *Signal Processing*, 99:215–249, 2014.

[21] Hidekata Hontani, Takamiti Matsuno, and Yoshihide Sawada. Robust nonrigid icp using outlier-sparsity regularization. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 174–181. IEEE, 2012.

[22] Pascal A Dufour, Hannan Abdillahi, Lala Ceklic, Ute Wolf-Schnurrbusch, and Jens Kowal. Pathology hinting as the combination of automatic segmentation with a statistical shape model. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 599–606. Springer, 2012.

[23] Hans Lamecker, Stefan Zachow, Antonia Wittmers, Britta Weber, H Hege, B Isholtz, and Michael Stiller. Automatic segmentation of mandibles in low-dose ct-data. *International Journal of Computer Assisted Radiology and Surgery*, 1:393, 2006.

[24] P. K. Saha, G. Liang, J. M. Elkins, A. Coimbra, L. T. Duong, D. S. Williams, and M. Sonka. A new osteophyte segmentation algorithm using the partial shape model and its applications to rabbit femur anterior cruciate ligament transection via micro-ct imaging. *IEEE Transactions on Biomedical Engineering*, 58(8):2212–2227, 2011.

[25] H Lamecker, S Zachow, H Hege, M Zockler, and H Haberl. Surgical treatment of craniosynostosis based on a statistical 3d-shape model: first clinical application. *International Journal of Computer Assisted Radiology and Surgery*, 1:253, 2006.

[26] Moritz Ehlke, Mark Heyland, Sven Märdian, Georg N Duda, and Stefan Zachow. Assessing the relative positioning of an osteosynthesis plate to the patient-specific femoral shape from plain 2d radiographs. 2015.

[27] Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery, 2017.

[28] Suhang You, Kerem Tezcan, Xiaoran Chen, and Ender Konukoglu. Unsupervised lesion detection via image restoration with a normative prior. In *International Conference on Medical Imaging with Deep Learning – Full Paper Track*, London, United Kingdom, 08–10 Jul 2019.

[29] Laura Estacio, Moritz Ehlke, Alexander Tack, Eveling Castro, Hans Lamecker, Rensso Mora, and Stefan Zachow. Unsupervised detection of disturbances in 2d radiographs. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 367–370. IEEE, 2021.

[30] Xiaobai Chen, Aleksey Golovinskiy, and Thomas Funkhouser. A benchmark for 3d mesh segmentation. *Acm transactions on graphics (tog)*, 28(3):1–12, 2009.

[31] D. Chetverikov, D. Svirko, D. Stepanov, and P. Krsek. The trimmed iterative closest point algorithm. In *Object recognition supported by user interaction for service robots*, volume 3, pages 545–548 vol.3, 2002.

[32] Brian Amberg, Sami Romdhani, and Thomas Vetter. Optimal step nonrigid icp algorithms for surface registration. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2007.

[33] Volker Roth. Kernel fisher discriminants for outlier detection. *Neural computation*, 18(4):942–960, 2006.

[34] Camila Gonzalez, Karol Gotkowski, Andreas Bucher, Ricarda Fischbach, Isabel Kaltenborn, and Anirban Mukhopadhyay. Detecting when pre-trained nnu-net models fail silently for covid-19 lung lesion segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 304–314. Springer, 2021.

[35] Jiacheng Cheng and Nuno Vasconcelos. Learning deep classifiers consistent with fine-grained novelty detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1664–1673, June 2021.

[36] Natasha Gelfand, Niloy J Mitra, Leonidas J Guibas, and Helmut Pottmann. Robust global registration. In *Symposium on geometry processing*, volume 2, page 5, 2005.

[37] Haili Chui and Anand Rangarajan. A new point matching algorithm for non-rigid registration. *Computer Vision and Image Understanding*, 89(2-3):114–141, 2003.

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

[39] Matthew Toews and Tal Arbel. A statistical parts-based model of anatomical variability. *IEEE Transactions on Medical Imaging*, 26(4):497–508, 2007.

[40] Shenzhi Wang, Liwei Wu, Lei Cui, and Yujun Shen. Glancing at the patch: Anomaly localization with global and local feature comparison. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 254–263, June 2021.

[41] Bernhard Egger. *Semantic morphable models*. PhD thesis, University_of_Basel, 2017.

[42] Sandro Schönborn, Bernhard Egger, Andreas Forster, and Thomas Vetter. Background modeling for generative image models. *Computer Vision and Image Understanding*, 136:117–127, 2015.

[43] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[44] Sergios Theodoridis and Konstantinos Koutroumbas. Pattern recognition. 2003. *Google Scholar Google Scholar Digital Library Digital Library*, 2009.

[45] Dana Rahbani, Andreas Morel-Forster, Dennis Madsen, Marcel Lüthi, and Thomas Vetter. Robust Registration of Statistical Shape Models for Unsupervised Pathology Annotation. In *Large-Scale Annotation of Biomedical Data and Expert Label Synthesis and Hardware Aware Learning for Medical Imaging and Computer Assisted Intervention*, pages 13–21. Springer, 2019.

[46] Tomas Pajdla and Luc Van Gool. Matching of 3-d curves using semi-differential invariants. In *Proceedings of IEEE International Conference on Computer Vision*, pages 390–395. IEEE, 1995.

[47] Dmitry Chetverikov, Dmitry Stepanov, and Pavel Krsek. Robust euclidean alignment of 3d point sets: the trimmed iterative closest point algorithm. *Image and Vision Computing*, 23(3):299 – 309, 2005.

[48] Sandro Schönborn, Bernhard Egger, Andreas Morel-Forster, and Thomas Vetter. Markov chain monte carlo for automated face image analysis. *International Journal of Computer Vision*, 123(2):160–183, 2017.

[49] Shachar Fleishman, Daniel Cohen-Or, and Cláudio T Silva. Robust moving least-squares fitting with sharp features. *ACM transactions on graphics (TOG)*, 24(3):544–552, 2005.

[50] Christian Payer, Darko Štern, Horst Bischof, and Martin Urschler. Integrating spatial configuration into heatmap regression based cnns for landmark localization. *Medical image analysis*, 54:207–219, 2019.

[51] Keenan Crane, Clarisse Weischedel, and Max Wardetzky. Geodesics in heat: A new approach to computing distance based on heat flow. *ACM Transactions on Graphics (TOG)*, 32(5):152, 2013.

[52] Fábio de Lima Bezerra and Jacques Wainer. A dynamic threshold algorithm for anomaly detection in logs of process aware systems. *Journal of Information and Data Management*, 2012.

[53] Rolf Adams and Leanne Bischof. Seeded region growing. *IEEE Transactions on pattern analysis and machine intelligence*, 16(6):641–647, 1994.

[54] Oliver Schall, Alexander Belyaev, and H-P Seidel. Robust filtering of noisy scattered point data. In *Proceedings Eurographics/IEEE VGTC Symposium Point-Based Graphics, 2005.*, pages 71–144. IEEE, 2005.

[55] Yutao Wang and Hsi-Yung Feng. Outlier detection for scanned point clouds using majority voting. *Computer-Aided Design*, 62:31–43, 2015.

[56] Marco Attene, Bianca Falcidieno, and Michela Spagnuolo. Hierarchical mesh segmentation based on fitting primitives. *The Visual Computer*, 22(3):181–193, 2006.

[57] Mark Pauly, Niloy J Mitra, and Leonidas J Guibas. Uncertainty and variability in point cloud surface data. *SPBG*, 4:77–84, 2004.

[58] David JC MacKay. Information-based objective functions for active data selection. *Neural computation*, 4(4):590–604, 1992.

[59] Robert B. Gramacy. *Surrogates: Gaussian Process Modeling, Design and Optimization for the Applied Sciences*. Chapman Hall/CRC, Boca Raton, Florida, 2020. http://bobby.gramacy.com/surrogates/.

[60] Kevin P. Murphy. *Probabilistic Machine Learning: An introduction*. MIT Press, 2021.

[61] Sambu Seo, Marko Wallat, Thore Graepel, and Klaus Obermayer. Gaussian process regression: Active data selection and test point rejection. In *Mustererkennung 2000*, pages 27–34. Springer, 2000.

[62] Shogo Iwazaki, Yu Inatsu, and Ichiro Takeuchi. Mean-variance analysis in bayesian optimization under uncertainty. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 973–981. PMLR, 13–15 Apr 2021.

[63] Donald R Jones, Matthias Schonlau, and William J Welch. Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13(4):455–492, 1998.

[64] Jonas Mockus, Vytautas Tiesis, and Antanas Zilinskas. The application of bayesian methods for seeking the extremum. *Towards global optimization*, 2(117-129):2, 1978.

[65] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, pages 586–606. International Society for Optics and Photonics, 1992.

[66] Dana Rahbani, Andreas Morel-Forster, Dennis Madsen, Jonathan Aellen, and Thomas Vetter. Sequential gaussian process regression for simultaneous pathology detection and shape reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 429–438. Springer, 2021.

[67] George EP Box and David R Cox. An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2):211–243, 1964.

[68] Jakob Raymaekers and Peter J Rousseeuw. Transforming variables to central normality. *Machine Learning*, pages 1–23, 2021.

[69] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[70] Dennis Madsen, Andreas Morel-Forster, Patrick Kahr, Dana Rahbani, Thomas Vetter, and Marcel Lüthi. A closest point proposal for mcmc-based probabilistic surface registration. *arXiv preprint arXiv:1907.01414*, 2019.

[71] Jianning Li and Jan Egger. *Towards the Automatization of Cranial Implant Design in Cranioplasty*. Springer, 2020.

[72] Nicholas Heller, Niranjan Sathianathen, Arveen Kalapara, Edward Walczak, Keenan Moore, Heather Kaluzniak, Joel Rosenberg, Paul Blake, Zachary Rengel, Makinna Oestreich, et al. The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. *arXiv preprint arXiv:1904.00445*, 2019.

[73] Michael Kistler, Serena Bonaretti, Marcel Pfahrer, Roman Niklaus, and Philippe Büchler. The virtual skeleton database: An open access repository for biomedical research and collaboration. *J Med Internet Res*, 15(11):e245, Nov 2013.

[74] Hans Lamecker, Dagmar Kainmueller, Stefan Zachow, et al. Automatic detection and classification of teeth in ct data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 609–616. Springer, 2012.

[75] Mike Biggs and Phil Marsden. Dental identification using 3d printed teeth following a mass fatality incident. *Journal of Forensic Radiology and Imaging*, 18:1–3, 2019.

[76] Jian Sun, Maks Ovsjanikov, and Leonidas Guibas. A concise and provably informative multi-scale signature based on heat diffusion. In

Computer graphics forum, volume 28, pages 1383–1392. Wiley Online Library, 2009.

[77] Vimal Chandran, Ghislain Maquer, Thomas Gerig, Philippe Zysset, and Mauricio Reyes. Supervised learning for bone shape and cortical thickness estimation from ct images for finite element analysis. *Medical Image Analysis*, 52:42–55, 2019.

[78] Joëlle Ackermann, Matthias Wieland, Armando Hoch, Reinhold Ganz, Jess G Snedeker, Martin R Oswald, Marc Pollefeys, Patrick O Zingg, Hooman Esfandiari, and Philipp Fürnstahl. A new approach to orthopedic surgery planning using deep reinforcement learning and simulation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 540–549. Springer, 2021.

[79] Raunak Dey and Yi Hong. Asc-net: Adversarial-based selective network for unsupervised anomaly segmentation. *arXiv preprint arXiv:2103.03664*, 2021.

[80] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4(1):1–13, 2017.

[81] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014.

[82] Ujjwal Baid, Satyam Ghodasara, Michel Bilello, Suyash Mohan, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021.

[83] Stefan Zachow, Alexander Steinmann, Thomas Hildebrandt, Rainer Weber, and Werner Heppt. Cfd simulation of nasal airflow: towards treatment planning for functional rhinosurgery. *International Journal of Computer Assisted Radiology and Surgery*, 1(7):165–167, 2006.

[84] Cornelius Reyneke, Xolisile Thusini, Tania Douglas, Thomas Vetter, and Tinashe Mutsvangwa. Construction and validation of image-based statistical shape and intensity models of bone. In *2018 3rd Biennial South African Biomedical Engineering Conference (SAIBMEC)*, pages 1–4. IEEE, 2018.

CV

Education

- (2017-2022) **PhD Computer Science**
  Graphics and Vision Lab, University of Basel, Switzerland
  Thesis: Outlier detection for shape model fitting
- (2015-2017) **MSc Biomedical Engineering**
  ETH Zurich, Switzerland
  Thesis: Modeling and simulation of leg dynamics after mechanical perturbations to the knee
- (2011-2015) **BEng Mechanical Engineering**
  American University of Beirut, Lebanon
  Thesis: Design and fabrication of an angular steering mobile robotic platform

Academic Experience

- (09/2021) Oral Presentation at International Conference of Robust Statistics ICORS21
- (09/2021) Poster Presentation at International Conference on Medical Image Computing and Computer Assisted Intervention MICCAI21
- (2019-2020) Mentee at ZOOM @ Novartis program for female researchers at the University of Basel
- (11/2019) Guest Speaker at Institute of Computer Graphics and Vision at TU Graz
- (07/2019) Poster Presentation at Computer Vision Summer School ICVSS2019

- (06/2019) Committee member at CVPR2019: Deep learning for geometric shape understanding
- (2017-Present) Instructor in BSc course Pattern Recognition and Supervisor in Machine Intelligence Seminar
- (2017-Present) MSc Thesis Supervisor

Publications

- Rahbani, Dana, et al. "Sequential Gaussian Process Regression for Simultaneous Pathology Detection and Shape Reconstruction." MICCAI. Springer, Cham, 2021.
- Rahbani, Dana, et al. "Robust Registration of Statistical Shape Models for Unsupervised Pathology Annotation." LABELS. Springer, Cham, 2019. 13-21.
- Madsen, Dennis, et al. "A closest point proposal for MCMC-based probabilistic surface registration." ECCV. Springer, Cham, 2020.
- I. Demir, et al."SkelNetOn 2019: Dataset and Challenge on Deep Learning for Geometric Shape Understanding." Proceedings of the CVPR Workshops. 2019.
- De Pieri, Enrico, et al. "Influence of muscle activity and co-contraction on hip contact forces prediction." In: XXVI Congress of the International Society of Biomechanics, 23-27 July 2017, Brisbane, Australia.
- Hasan, Anwarul, et al. "Engineered biomaterials to enhance stem cell-based cardiac tissue engineering and therapy." Macromolecular bioscience 16.7 (2016): 958-977.

Skills

- Languages: Arabic (fluent), English (fluent), German (Goethe Exam B2), French (intermediate)
- Software: Scala, Python, MATLAB, Mathematica, R, LaTeX, Microsoft Office