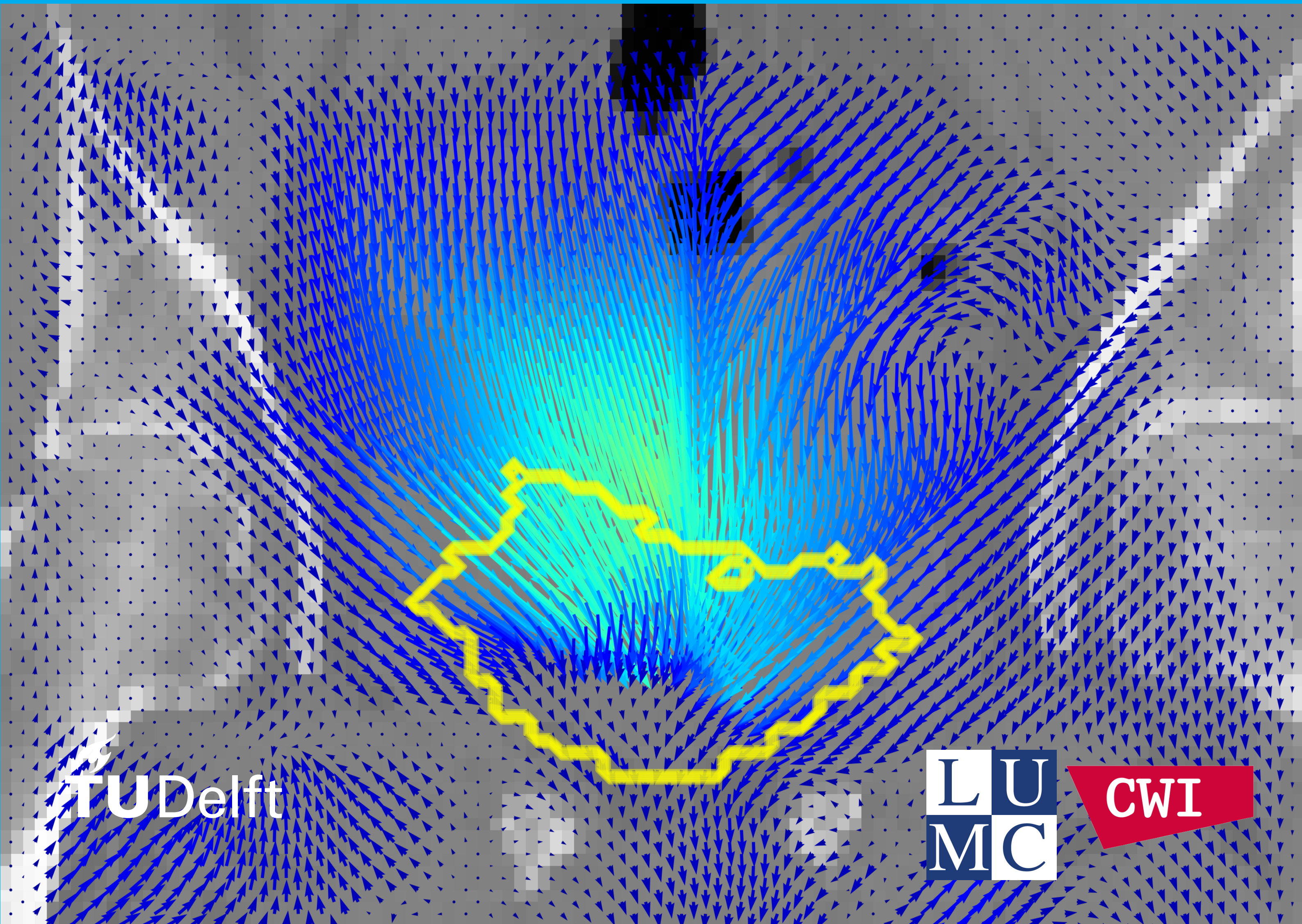
On the Hybridization of Automated Landmark Detection with Physics-Based Multi-Objective Deformable Image Registration with an Application to Radiotherapy

C.J. Wever

MSc Thesis



On the Hybridization of Automated Landmark Detection with Physics-Based Multi-Objective Deformable Image Registration with an Application to Radiotherapy

by

C.J. Wever

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Friday June 28, 2024 at 10:00 AM.

Student number:	4885619	
Project duration:	November 13, 2023 – June 28, 2024	
Thesis committee:	Prof. dr. P.A.N. Bosman,	TU Delft, Algorithmics, supervisor
	Dr. T. Alderliesten,	LUMC, Radiation Oncology, supervisor
	Dr. A. Panichella,	TU Delft, Software Engineering
	Ir. G. Andreadis,	LUMC, daily supervisor
	Ir. C.J. Rodriguez,	LUMC, daily supervisor

Abstract

Deformable Image Registration (DIR) is a process in which the point-to-point correspondence between two or more medical images is estimated. This could allow spatial data to be transferred between these images, easing the work of practitioners in the field of radiation oncology. There exist many DIR approaches. These are not yet applicable in all clinical settings, however, as DIR is a complex problem, making it difficult to evaluate. This is why often multiple objectives are considered when evaluating a deformed image. In this thesis, we combine two separate multi-objective DIR approaches into hybrids while evaluating their performances, with the aim to improve upon the individual approaches, thus reducing the gap toward clinical practice. In these hybrids, we also use an approach called Automated Landmark Detection which automatically detects corresponding landmarks between the source and target images.

The first DIR approach, the Digital Phantom, creates a biomechanical model of the patient and uses an Evolutionary Algorithm (EA) to optimize parameters of a Finite Element Method (FEM)-based simulation. The second DIR approach, MOREA, creates a dual-dynamic biomechanical transformation model of the patient, using an EA to optimize its parameters.

In the first hybrid approach, called the Landmark Guided Digital Phantom (LGDP), the Automated Landmark Detection is used to automatically find corresponding landmarks, which are then used as a guidance objective in the Digital Phantom. The second hybrid is called DP-MOREA, where the output of the Digital Phantom is used to initialize deformations in the MOREA approach. The final hybrid is called LGDP-MOREA. This hybrid combines all three components, using the generated landmarks to guide the Digital Phantom and initializing MOREA with the result. We analyze the effect of these hybrids on the quality of the final registration, both quantitatively and qualitatively. Additionally, we compare these hybrids with the components they are composed of.

We test the performance of the approaches and hybrids on pelvic CT scans of three cervical cancer patients with large deformations. We observe that the individual Digital Phantom component outperforms the LGDP hybrid both quantitatively and qualitatively on these three patients. In addition, we observe that initializing MOREA with the output of the Digital Phantom and LGDP improves on how fast the optimization converges, that it improved the end results quantitatively and that it also has an effect on the resulting Deformation Vector Fields (DVFs). We conclude that more work is required to investigate these effects and their clinical relevance.

Preface

I started this thesis with somewhat of a deep dive. Coming into the medical world with my background as a Computer Scientist took some adjusting. It was a refreshing way to look at how the knowledge I have obtained during the last 6 years can be used to help people, however, and was incredibly rewarding. Working on this thesis project did pose several challenges that I would not have overcome without the help of several people.

First of all, I would like to thank my thesis committee for all the guidance during the project. To Peter and Tanja, thank you for the detailed feedback and insights into the results I obtained and for providing future directions I did not always see. To Cedric and Georgios, thank you for the meetings, feedback, and detailed brainstorms that we shared. You helped me come through the difficult gap between the Computer Science world and the medical world, and ensured I really understood everything, even toward the end. I appreciate that I could always send you a message with questions, no matter how small or at what time, and rely upon your feedback.

Finally, I would like to thank my family and friends for their support during my entire period at the university. I could always talk to you about anything, which always helped me get back on track. Without you, I would not have been able to make it this far.

C.J. Wever
Delft, June 2024

Contents

1

Introduction

In the field of radiation oncology, tumor tissue is treated with radiation. Treatment is typically delivered in multiple smaller fractions of radiation. Before each fraction is delivered, medical images are taken to establish the current location and shape of the tumor and surrounding healthy tissue. This ensures that the tumor can be targeted with high precision and surrounding tissue can be spared as much as possible. These images can be taken with various modalities, such as Computed Tomography (CT) or Magnetic Resonance Imaging (MRI). Before the start of treatment, a plan is made determining how much radiation dose each tissue location should receive across fractions. Most tissue can be subject to large deformation, however. In the pelvic region, for example, the uterus has been observed to move up to 4.8cm in between fractions (Taylor and Powell, 2008). This means that medical images of the same patient taken at different time points can appear different, even though the same tissue is captured. Furthermore, the content being imaged can differ between images. In one image, a surgical device might be present, while in the other, the device may have been removed. The insertion of such a surgical device not only causes content mismatch between the images, but also deforms the surrounding tissue.

Such large deformations are difficult to model, making it challenging to find the exact point-to-point correspondence between two images. The process of finding this mapping is called image registration. If found reliably, such a mapping could make it possible to transfer spatial information such as dose information between images (Oh and Kim, 2017). Without the possibility of transferring such spatial information, practitioners have to resort to other, potentially less accurate manual estimates.

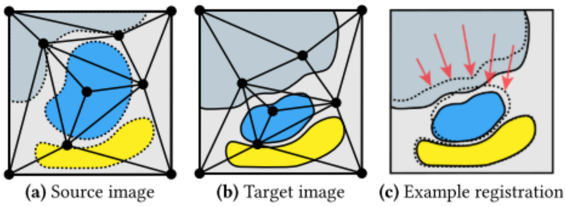

(a) Source image (b) Target image (c) Example registration

Figure 1.1: An illustration of Deformable Image Registration between two images (Andreadis et al., 2023).

In current clinical practice, a baseline mapping is established by translation and rotation of the images, typically aligning the rigid bony anatomy of the patient. This alignment is called rigid registration. Rigid registrations are, however, not able to model local deformations, as they only define linear, global transformations. To be able to model local deformations, a Deformable Image Registration (DIR) method is required. A DIR method is tasked with estimating “spatial correspondences between different image acquisitions” (Sotiras et al., 2013). This typically entails finding a mapping of the voxels (three-dimensional image pixels) of a so-called *source* image I_s to a second, *target* image I_t. This mapping makes it possible to transfer spatial data from one image to the other. It is estimated through a transformation T such that $T(I_s) \approx I_t$. Finding T is the goal of any DIR method, and often achieved by optimizing a certain optimization function. Figure 1.1 contains an example of what DIR can look like. Given the found correspondence, practitioners could adapt the treatment plan for a more precise delivery. Although several commercial DIR methods have been introduced, their use in clinical practice remains limited.

A potential cause for the lack of clinical adoption is caused by the inherently ill-posedness of the problem. Since the ground truth deformation is not known, it is difficult to evaluate the quality of a found deformation. Therefore, there is not one single objective to optimize which could then produce the "optimal" registration. There are multiple, conflicting objectives instead, that can all be used to evaluate a registration from different perspectives. The conflict between these objectives makes it inherently difficult to combine them in a single objective with weights assigned to them, as the weights needed can be different for every patient (Pirpinia et al., 2017). This calls for a multi-objective approach to the DIR problem.

1.1. Multi-Objective Nature

As mentioned, a registration can be evaluated using different objectives, each of which uses different types of metrics. These metrics can be categorized into three general types (Andreadis et al., 2023): *image intensity similarity*, *contour similarity*, and *deformation magnitude*. In this thesis, we will also consider a fourth type, which we call *landmark correspondence*.

The *image intensity similarity* objective builds upon the fact that each image is represented by a voxel grid of different intensity values. Comparing image intensities between locations of the deformed source image and the target image can therefore be used as one indicator of the accuracy and/or quality of the estimated transformation. Visually, this means that we want these images to look as similar as possible.

The *contour similarity* objective considers relevant organ contours. These can either be manually annotated by field experts or automatically generated by an algorithm. Quantifying the similarity between corresponding organ contours can indicate how well the transformed source image and the target image correspond in terms of clinical metadata.

The *deformation magnitude* objective measures the energy required to perform a transformation. The energy required for a certain transformation can depend on factors such as the elasticity of deformed tissues and the homogeneity of the transformation. A high amount of required energy implies that the transformation likely contains large deformations and is less homogeneous, while a small amount of energy implies a more homogeneous transformation.

Finally, the *landmark correspondence* objective uses a function to measure the distance between corresponding landmarks in between the images. A corresponding landmark is a landmark voxel location in the source image which corresponds to a landmark voxel location in the target image. This correspondence can be used as guidance information, as the found transformation should translate source landmarks close to their corresponding target landmarks. These landmarks can, like the organ contours, either be manually annotated by field experts or automatically generated.

These different objectives form a tradeoff. An example is the regularizing function that the deformation magnitude has. Consider the case where the deformation magnitude is minimized, while the contour similarity is maximized. Here, having a better contour similarity score could require a larger, less homogeneous deformation, thus resulting in a worse deformation magnitude. Deformation magnitude is, in this case, a regularizing objective. There can also be other forms of tradeoffs, such as inconsistent or incomplete data. When DIR problems are viewed from this perspective, it is possible to obtain a set of multiple high-quality registrations that effectively represent the tradeoff between objectives, instead of only one registration. This removes the need to perform a scalarization, where weights need to be assigned to each objective (Alderliesten et al., 2015).

1.2. Contribution

Even though there are many potential use cases for DIR, it is still not widely used in some clinical applications (Rigaud et al., 2019). We highlight two DIR approaches that form the core of this research. They each model the problem differently. Additionally, we highlight another approach that can serve as input to DIR approaches.

One approach, the Digital Phantom (Rodriguez et al., 2023), leverages a Finite Element Method (FEM) based simulation to solve the problem. The parameters of a simulation framework are optimized with the contour similarity and deformation magnitude objectives, while the space of simulation parameters is explored using an Evolutionary Algorithm (EA). The advantage of this approach is that it models the problem from a biomechanical perspective, where the external and internal forces on tissue are simulated and optimized. Additionally, it can contain a preprocessing step allowing to simulate a large deformation and content mismatch due to insertion of a surgical device before optimization. The main disadvantage of this approach is the indirect nature of the problem being optimized. The parameters that the EA optimizes represent global forces on the modeled tissue, which only indirectly influence the transformation of the model. As a result, deformations can be less fine-grained, as the deformed model may potentially not be able to reach the desired state, exactly.

Another approach, MOREA (Andreadis et al., 2023), optimizes deformation with an EA, using the image intensity, contour similarity, and deformation magnitude objectives. In this approach, a dual-dynamic mesh transformation model is deformed using different parameter settings. One advantage of MOREA is that it models tissue-specific elasticities, allowing the deformation magnitude objective to be estimated more accurately. This could possibly lead to a more fine-grained optimization of the deformation, as well. This is related to the increased degrees of freedom that MOREA's model has, meaning that a large number of parameters is optimized. Additionally, these parameters directly influence the mesh, instead of the indirect influence of the Digital Phantom. The disadvantage of MOREA is that the degrees of freedom create a much larger search space, demanding more computational resources. Moreover, the inverse nature of MOREA's model simulation can be less realistic than the forward nature of the simulations within the Digital Phantom method.

The third approach, the Automated Landmark Detection approach (Grewal et al., 2020), uses deep learning and the intensity values of a scan to find corresponding landmark locations between two images. This correspondence contains information on where certain points in the images may need to be translated to in a transformation. The advantage of this approach is that it can automatically create correspondences between different images without any supervision. However, this is not a registration method on its own, and can only indirectly inform a registration method, which then can estimate a transformation based on this correspondence information. Additionally, it is an unsupervised approach, meaning that the deformations from which it learns might not reflect real-world deformations. This could result in less accurate landmark correspondences.

These three approaches will be examined in more detail in Section 2. A hybrid approach combining these different approaches may improve upon the individual approaches because each has the different abovementioned strengths and weaknesses that might be mitigated in this combination. Therefore, in this thesis, different hybrid combinations of these approaches will be designed, implemented, and tested.

1.2.1. Landmark Guided Digital Phantom

The first hybrid that we will consider is a link between the Automated Landmarks approach and the Digital Phantom DIR approach, where we will use the landmarks as a guidance objective in the EA of the Digital Phantom. The pipeline of this hybrid can be viewed in Figure 1.2 and will be called the *Landmark Guided Digital Phantom* (LGDP). This hybrid is motivated by the hypothesis that these landmarks could provide more fine-grained guidance information to the Digital Phantom, possibly also resulting in a more fine-grained result of the Digital Phantom. Since the landmarks are only used indirectly in optimization, this could

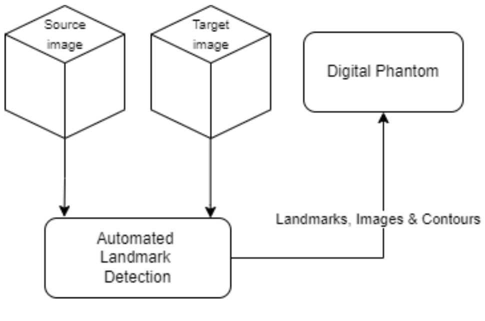


Figure 1.2: A global overview of the Landmark Guided Digital Phantom hybrid.

also decrease the potential impact of any inaccurate generated landmarks.

1.2.2. DP-MOREA

The second hybrid is a combination of the Digital Phantom and MOREA, where the output of the Digital Phantom is used in the initialization of MOREA, before its optimization loop of the deformation. This hybrid can be seen in Figure 1.3 and is called *DP-MOREA*. It leaves the global exploration of the problem to the Digital Phantom and leverages MOREA to locally further develop its generated solutions. This allows us to exploit the large number of degrees of freedom of MOREA, compared to the Digital Phantom, and to fine-tune these solutions, while reducing the search space for MOREA by leveraging the global search capability of the Digital Phantom as an initialization step.

1.2.3. LGDP-MOREA

The final hybrid is a combination of the Landmark Guided Digital Phantom and DP-MOREA, where the output of the Landmark Guided Digital Phantom will be used to initialize MOREA. This hybrid can be viewed in Figure 1.3, and we will refer to it as *LGDP-MOREA*. LGDP-MOREA combines the guidance information of the landmarks, the global search capabilities of the Digital Phantom and the local search capabilities of MOREA, possibly finding more accurate registrations than the individual approaches.

1.3. Research Questions

To determine the performance of each hybrid, three main research questions have been formulated:

1. What is the quality of the transformations of the Landmark Guided Digital Phantom, and how fast does it converge to these transformations, when compared with the Digital Phantom?
2. What is the quality of the transformations of DP-MOREA, and how fast does it converge to these transformations, when compared with its individual components?
3. What is the quality of the transformations of LGDP-MOREA, and how fast does it converge to these transformations, when compared with its individual components?

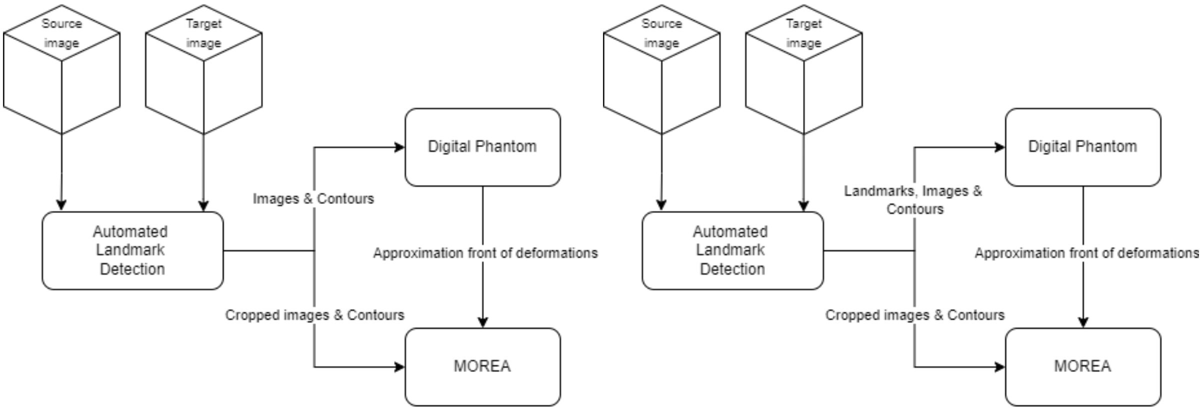


Figure 1.3: A global overview of the DP-MOREA (left) and LGDP-MOREA (right) hybrids.

2

Background

In this section, we introduce key concepts in DIR. Furthermore, we review literature relevant to the contributions of this thesis.

2.1. Deformable Image Registration

As mentioned in Chapter 1, the goal of any DIR method is to find a mapping between a *source* and *target* image, I_s and I_t, respectively. The source image is the image from which to estimate the transformation and ultimately the image undergoing the deformation. The target image is the image we are trying to approximate when applying the estimated transformation to the source image. The estimated transformation is usually expressed in the form of a Deformation Vector Field (DVF), which is a way of representing a transformation T on all voxel positions $P \in I$ by assigning a vector to each P. Each vector is determined by the difference between corresponding points P and $T(P)$, thus representing where P is moved by T. This representation allows to easily re-apply the same transformation between multiple images.

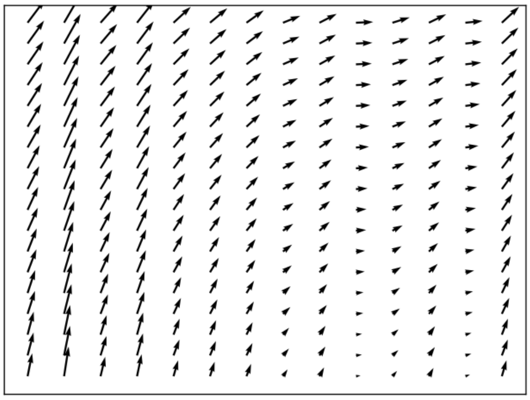


Figure 2.1: An example of a Deformation Vector Field.

A DVF can be defined in two ways: forward and inverse. A forward DVF describes the transformation $T(I_s)$ to I_t, while an Inverse DVF describes the transformation $T^{-1}(I_t)$ to I_s. A DVF whose transformation describes the true deformation between image spaces I_s and I_t is thus forward when $T(I_s) = I_t$ and inverse when $I_s = T(I_t)$. Figure 2.1 shows a visual example of a DVF. We often cannot analytically exactly obtain $T^{-}1$ from T due to the nonlinearity of the deformation, except in the case where the DVF is ensured to be *inverse consistent*. Ensuring inverse consistency means that, if we estimate a deformation using transformations $T_f(I_s) \approx I_t$ and $T_i(I_t) \approx I_s$, we ensure that $T_f^{-1} = T_i$ and $T_f = T_i^{-1}$.

Performing DIR generally requires choosing an implementation for three major components: A transformation model, a function determining the quality of a registration, and an optimization method (Sotiras et al., 2013).

2.1.1. Transformation Model

The transformation model determines the internal representation of the problem and potential solutions but also establishes and limits the deformations that can be performed. One could, for example, model a deformation by creating a mesh model and moving its nodes. Another example is to place grid points on the images and model the deformation by changing the point coordinates and parameterizing the deformation in that way (Weistrand and Svensson, 2015).

Many different (geometric) transformation models exist (Sotiras et al., 2013). They can be broadly categorized into models based on physical models, on interpolation and approximation theory, and those based on knowledge. There are also many subcategories within these categories, but in this research, we are mainly interested in the elastic body model and the knowledge-based model subcategories. In the elastic body model category, the problem is generally modeled as a biomechanical mesh upon which forces are applied. These forces result in a deformation in the mesh, with which the target deformation can be established. The Digital Phantom approach and MOREA are examples of such a model. Knowledge-based models attempt to perform registration based on previously obtained knowledge. This can, for example, be done by using deep-learning to predict a deformation in a supervised or unsupervised manner.

Choosing a model comprises a tradeoff between the resolution of the transformation and computational efficiency. Every unique model requires a different number of parameters to optimize, related to the degrees of freedom needed for the transformation. Since the true transformation can be non-linear and difficult to describe, having a model with more degrees of freedom could result in a more accurate and/or descriptive transformation. This does, however, also increase the complexity of the model, creating a larger problem space to explore.

2.1.2. Objective Functions

Defining an objective function that quantifies the quality of the solution is of high importance to the optimization. An objective function has to represent the objective to be achieved as closely as possible to ensure the solution is of a desired quality. One cannot, for example, use the deformation energy to determine how well contours between $T(I_s)$ and I_t overlap. As mentioned in Chapter 1, we identify four types of metrics: image intensity similarity, contour similarity, deformation magnitude, and landmark correspondence. Several existing and commonly used metrics will be discussed below.

Dice Coefficient

One way to evaluate the contour similarity objective is by using the Dice coefficient (Dice, 1945). It computes the fraction of volumetric overlap between two binary masks (representing organs or objects being imaged). Consider Figure 2.2 for a 2-dimensional example of contour overlap between two areas, A and B.

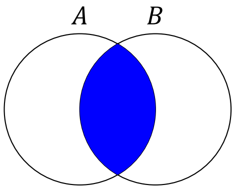


Figure 2.2: A 2-dimensional example of the overlap between areas A and B.

Equation 2.1 shows how this coefficient is computed between two sets A and B. This formula yields a value of 0 for non-overlapping sets and a value of 1 for $A = B$. Therefore, the more overlapping two sets are, the higher the resulting Dice score will be. For contour similarity, the overlap between two organ segmentations can be calculated by deriving voxel sets from volumes A and B (De Vos et al., 2019).

$$Dice = \frac{2 * |A \wedge B|}{|A| + |B|} \tag{2.1}$$

Contours as Guidance Objective

MOREA uses another strategy to leverage organ contours. It calculates and sums the closest distances between two contours by sampling from Euclidean distance maps at locations close to either contour. To this end, two distance maps are used for each organ, one on the source and one on the target image.

Sum of Squared Differences

A way to compute the image intensity similarity is by computing the sum of squared differences between the intensities of the image voxels. This calculation can either be applied to the images as a whole, or to a selected subset of voxels. Intuitively, the smaller this summed difference, the more similar the compared images are.

Deformation Energy

Optimizing (minimizing) deformation energy as an objective is a way to evaluate the deformation magnitude. This metric represents the amount of energy required to perform the modeled deformation. Minimizing this deformation means regularizing the solutions, as it penalizes less homogeneous deformations. MOREA calculates the deformation energy by aggregating the difference in tetrahedron edge lengths multiplied with tissue-specific elasticity constants, for example.

To compute this value, each tetrahedron in the meshes that are generated by MOREA is assigned its own unique elasticity constant. This is based on the fact that different tissues have different elasticities, thus allowing for a more realistic model. The deformation magnitude is then computed by solving, for each edge e_s and e_t of each tetrahedron $\delta \in \Delta$ (where each tetrahedron has 10 edges due to spoke edges):

$$f_{magnitude} = \frac{1}{10|\Delta|} \sum_{\delta \in \Delta} \left[\sum_{(e_s, e_t) \in E_\delta} c_\delta (||e_s|| - ||e_t||)^2 \right] \tag{2.2}$$

Landmarks as Guidance Objective

Another possible strategy to evaluate the similarity between $T(I_s)$ and I_t is by leveraging the existence of corresponding landmarks. The distance between two corresponding landmarks on $T(I_s)$ and I_t indicates how far the transformation is from obtaining the same landmark positions on $T(I_s)$ and I_t. This distance can be computed using the Euclidean Distance aggregated over all landmarks, for example.

Multiple objectives

It is important to choose how many and what metrics to consider during optimization. It is possible to model the problem from a single-objective perspective, where either only one metric is used or multiple metrics are scaled together. Considering only one objective neglects the existence of other, possibly important objectives, however. Next to this, each objective function has its inaccuracies and can conflict with other objective functions. Additionally, scaling multiple metrics together requires to set the metric weights *a priori*. This calls for a multi-objective optimization approach.

2.1.3. Multi-Objective Deformable Image Registration

Instead of obtaining a single solution, a multi-objective approach finds a set of solutions which is called the *Approximation Set*. This set consists of solutions that are *non-dominated* by any other solution. A solution A is considered dominated by a solution B, denoted $B \succ A$, when for all objective values of B, there is at least one value strictly better than A, and none worse. We consider cases where the metrics are typically *minimized*, meaning that better in this case is lower. The set of non-dominated solutions represents a tradeoff between all different objective values. Figure 2.3 provides an example of a non-dominated front along the space of possible objective metric values. Being able to choose between different solutions from the non-dominated set can be beneficial to users, as this allows an *a posteriori* exploration of the tradeoff space.

2.1.4. Optimization Method

The optimization method determines how the variables of the transformation model are optimized to find high-quality solutions according to the objective function(s) (Sotiras et al., 2013). Choosing a well-fitting optimization method for the problem is important for the quality of the final solution. The optimizer should support multiple objectives. Optimization can be done using continuous or discrete variables, depending on the transformation model. One example of a continuous optimization method is gradient-based optimization, where generally the search direction is taken in parameter space that decreases (improves) the objective function. EAs are another category of optimization methods, which have been shown to be well-suited for multi-objective problems (Pirpinia et al., 2017).

2.2. Evolutionary Algorithms

An EA is a population-based approach to optimization which is very well suited to Multi-Objective optimization, as an EA can be leveraged to simultaneously find non-dominated solutions (Deb, 2011). This allows to explore the objective space of the problem automatically, resulting in a non-dominated front

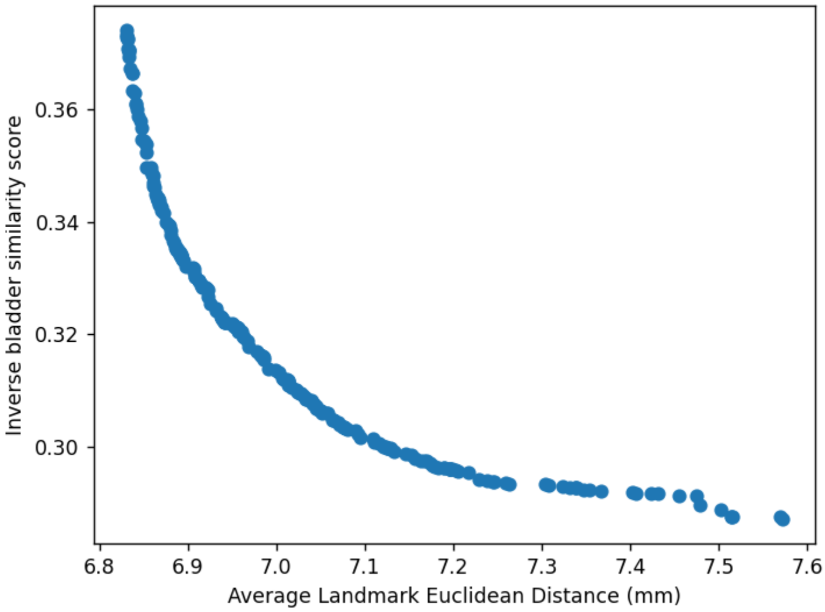


Figure 2.3: An example of a non-dominated front along the objective space.

of solutions. An EA iteratively optimizes a problem by dividing the search into generations, where the existing population of solutions is simultaneously evaluated, recombined, and mutated. A part of this newly generated population is selected and forms the population for the next generation. Typically, this generated population is based on the best-evaluated solutions from the current population. This means that, with each generation, the EA attempts to approximate the (unknown) optimal Pareto Front in terms of objective metric values. In the next two subsections, we describe two EAs that we use in this research.

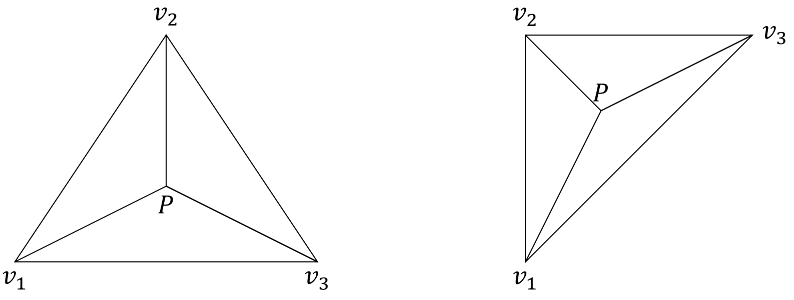
2.2.1. RVEA

RVEA (Reference Vector Guided Evolutionary Algorithm) (Cheng et al., 2016) is a multi-objective EA. In its optimization process, it generates uniformly distributed *reference vectors*, which are unit vectors representing objective values in an *n*-dimensional space, where *n* is the number of objective values. For every generation, offspring is created from the parent population using simulated binary crossover (Deb, Agrawal, et al., 1995), where parents are randomly selected from a subset of the current population. After this step, the parent and child populations are combined. Each solution in this new population is evaluated and mapped to a reference vector, effectively creating subpopulations. Since the reference vectors represent objective values, a solution will also be represented by its objective values. Finally, from each subpopulation, one solution is selected that will fulfill the role of parent for the next generation. This solution is the one that has a minimal value for the *Angle-Penalized Distance*, which tries to balance convergence and diversity with respect to the assigned reference vector. After this step, the reference vectors are adapted to the ranges of the objective values to further promote diversity in solutions and a new generation starts.

2.2.2. MO-RV-GOMEA

MO-RV-GOMEA (Multi-Objective Real-Valued Gene-pool Optimal Mixing Evolutionary Algorithm) (Bouter et al., 2017) is an EA which leverages Family-of-Subsets (FOS) linkage models to estimate relations between problem variables. In this type of linkage model, subsets of problem variables are defined, where each subset ideally contains variables that have some dependency on one another. The model can either be defined beforehand, in the case of Gray-Box Optimization (GBO) where a problem structure is known, or during optimization from candidate solutions in each generation.

Non-dominated solutions are contained in an elitist archive. This archive preserves solutions that are either non-dominated by other solutions in the elitist archive, or dominate at least one other solution in



(a) Source triangle (b) Target triangle

Figure 2.4: Example of the movement of point P on a triangle when its vertices are moved. In this instance, $v1$ remains stationary.

the archive. For each generation, a set of non-dominated solutions selected from the current population is clustered into q equally sized, non-overlapping clusters. For each cluster, a FOS linkage model is used to perform variation by sampling the multivariate Gaussian distribution of the model for each set in this linkage model. For each linkage set, a change is only accepted if the resulting solution dominates the parent solution, or if it is non-dominated by the current elitist archive.

2.3. Barycentric Coordinates for Tetrahedra

Barycentric coordinates form a representation of a point inside a triangle (Floater, 2015). In a 2-dimensional space and for a triangle consisting of vertices v_1, v_2, and v_3, a point P is defined with respect to the coordinates of this triangle as

$$P = \phi_1 v_1 + \phi_2 v_2 + \phi_3 v_3 \tag{2.3}$$

In this case, $\phi_i = A_i/A$, where A is the area of the triangle, and areas A_i are the areas of the formed sub-triangles by adding P to the triangle. It logically follows that $\sum_i A_i = A$. Dividing this equation by A and substituting $\phi_i = A_i/A$, we obtain

$$\sum_i \phi_i = 1 \tag{2.4}$$

Using this information, the barycentric coordinates ϕ_i of $P = (P_x, P_y)$ in 2D can be calculated by solving a system of linear equations:

$$\begin{cases} \phi_1 v_1^x + \phi_2 v_2^x + \phi_3 v_3^x = P_x \\ \phi_1 v_1^y + \phi_2 v_2^y + \phi_3 v_3^y = P_y \\ \phi_1 + \phi_2 + \phi_3 = 1 \end{cases} \tag{2.5}$$

Figure 2.4 shows an example of a case where the location of point P moves due to the movement of the triangle vertices. The ratio of sub-areas formed by connecting P with the triangle vertices and the triangle volume remains the same.

In this research, we want to apply points in 3D space on tetrahedra, however. Extending from a triangle to a tetrahedron essentially means that we add a point vector to our original triangle while extending to 3D means extending the size of our vectors from 2 to 3. This implies that all original areas A_i become volumes, but since A is now a volume as well, Equation 2.4 does not change. Based on this information, we can extend Equation 2.3 and 2.4 to

$$\begin{gathered} P = \phi_1 v_1 + \phi_2 v_2 + \phi_3 v_3 + \phi_4 v_4 \\ s.t. \sum_i \phi_i = 1 \end{gathered} \tag{2.6}$$

Using this extension, we can now also extend Equation 2.5 to:

$$\begin{cases} \phi_1 v_1^x + \phi_2 v_2^x + \phi_3 v_3^x + \phi_4 v_4^x = P_x \\ \phi_1 v_1^y + \phi_2 v_2^y + \phi_3 v_3^y + \phi_4 v_4^x = P_y \\ \phi_1 v_1^z + \phi_2 v_2^z + \phi_3 v_3^z + \phi_4 v_4^x = P_z \\ \phi_1 + \phi_2 + \phi_3 + \phi_4 = 1 \end{cases} \tag{2.7}$$

Solving this linear system provides us with the barycentric coordinates of P. We can use this when computing the movement of P when its tetrahedron is deformed. This can be relevant for computations in mesh models and for the generation of DVFs. This will be further discussed in Chapter 3.

2.4. Digital Phantom

The Digital Phantom approach (Rodriguez et al., 2023) optimizes parameters for Finite Element Method (FEM) simulations run on the pelvic region, using the SOFA framework[1]. It uses a multi-objective estimation-of-distribution EA (MAMaLGaM, Bosman, 2010) and bases its simulations on a detailed biomechanical model to ensure all deformations are physics-based. An example of a biomechanical model generated by the Digital Phantom can be viewed in Figure 2.5.

To acquire realistic simulations, the model and the exerted forces on them should be tuned on a patient-to-patient basis. This tuning problem is considered bi-objective, due to the existence of multiple organs in the region: the bladder and the vagina-cervix-uterus (considered and modelled jointly). The two objectives represent the quality of each organ deformation, respectively.

The work is a proof-of-concept showing the possibilities of using FEM-based simulations for DIR. It is important to note that the version of the Digital Phantom described in the paper differs from the version used in this work. Details on this new version can be found in Chapter 4.

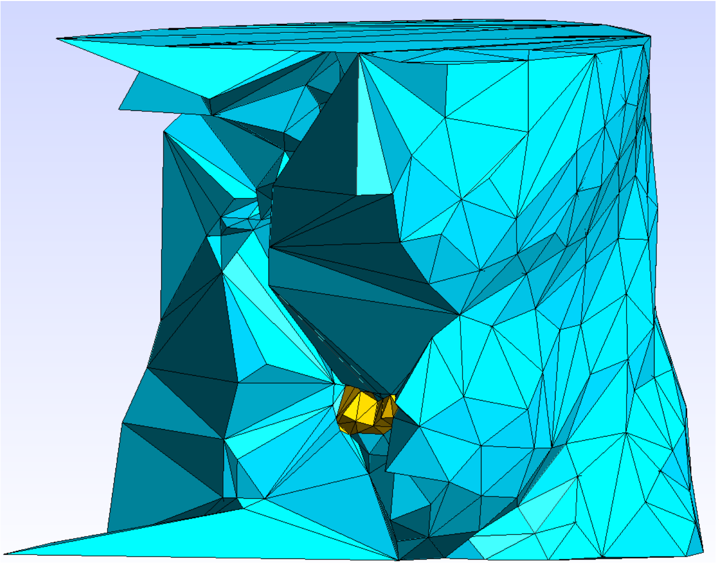


Figure 2.5: A clipped example of the tetrahedral mesh generated by the Digital Phantom approach, from a frontal perspective. Tetrahedra belonging to the body are indicated in blue, while those belonging to the bladder are indicated in yellow.

2.4.1. Preprocessing

The algorithm starts by transforming organ contours into a tetrahedral mesh representation. After this step, a rigid registration step is applied to the meshes to ensure the bony anatomy of the source and target meshes is aligned. When this is done, the initial deformation of the vagina-uterus, caused by

---

[1] https://www.sofa-framework.org/

inserting the applicator used in brachytherapy, is simulated. After this, the main optimization loop begins.

2.4.2. FEM Modeling

In the FEM-based simulation, a system of differential equations is solved for a three-dimensional space. The space is divided into smaller finite elements, in this case tetrahedra. Each element has its own mechanical deformation laws with associated material properties, loading forces, and constraints (fixations). The core intuition of this simulation is that constraints and functions that apply to large structures can also be applied to smaller parts of these structures, which allows for a model that can simulate local occurrences (Huebner et al., 2001). All material properties (elasticity of the uterus tissue), forces (pressure in the bladder), and fixations should be fully defined to be able to run a FEM simulation. The material properties and forces are unknown beforehand and show large variability between patients. These simulation parameters are therefore bi-objectively tuned on a patient-to-patient basis.

2.4.3. Optimization parameters

Each solution represents a set of simulation parameters for the FEM-based simulation software SOFA. These include a constant external force on the vagina-uterus mesh, which is represented by three parameters, and the tissue elasticity, represented by one parameter. Additionally, in practice, the applicator which is inserted into the vagina is fixed with pads. This is also modeled by placing four objects inside the vagina cavity, of which the diameters are represented by four parameters. Finally, the internal forces of the bladder are modeled by three parameters and the target bladder volume by one. This results in a total of twelve simulation parameters that need to be optimized.

2.4.4. Main Optimization Loop

Optimization is performed using MAMaLGaM. For every solution, a simulation is performed, where the parameters to optimize are used. The quality of each solution is evaluated by determining the quality of the result of the simulation. The two objectives used during optimization are the inverse Dice score of the bladder contours and the inverse Dice score of the vagina-cervix-uterus contours of $T(I_s)$ and I_t.

2.4.5. Optimization Results

The output of the algorithm is a set of solutions, each consisting of material properties and external forces, which can be used to acquire deformed organ meshes of the bladder and the vagina-cervix-uterus tetrahedral meshes. Each deformed mesh then represents a physics-based deformation when considered in relation to the original mesh configuration.

2.5. MOREA

MOREA is another DIR approach using a biomechanical mesh. It leverages MO-RV-GOMEA to optimize three objectives: image intensity similarity, contour similarity, and deformation magnitude. MOREA leverages the Graphics Processing Unit (GPU) to accelerate the direct optimization of its mesh model. The version of MOREA described in the paper also differs slightly from the version used in this work. Again, we provide details on this new version in Chapter 4.

2.5.1. Modeling

An important modeling aspect of MOREA is its usage of a *dual-dynamic* mesh transformation model. Here, two meshes are initialized with the same topology: one on the source image, and one on the target image. The mesh point positions of both are manipulated during optimization. This allows the algorithm to more easily take disappearing structures and large deformations into account.

Three objectives are used in optimization: deformation magnitude, image intensity similarity, and contour similarity. They are all calculated in the way described in 2.1.2. It is important to note that for the image intensity similarity and contour similarity objectives, a randomly sampled subset of coordinates in the images is used, instead of all discrete voxel positions in image space.

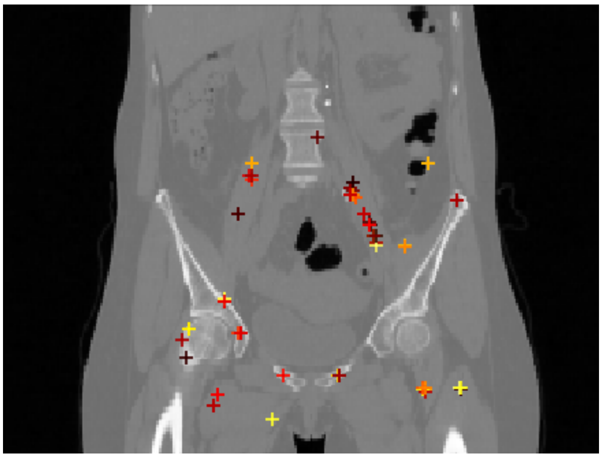


Figure 2.6: An example of generated landmarks on a CT scan, as seen from the coronal (front) view.

2.5.2. Preprocessing

The images and their contours are preprocessed by MOREA to generate the guidance points and the source mesh point coordinates. These point coordinates are transformed into a mesh by applying a Delaunay tetrahedralization. This source mesh is copied to the target mesh. These mesh points are also used in this preprocessing to determine the FOS elements that can be optimized in parallel, on the GPU. To ensure diversity in the initial solution meshes, small deviations are applied to both source and target mesh point coordinates.

2.5.3. Main Optimization Loop

MOREA optimizes the problem with an extension of MO-RV-GOMEA. To achieve this, the linkage models for each cluster are based on the topological proximity of variables. Additionally, for each new solution that is created, two techniques are applied to ensure quality: Repairing and adaptive steering. The repairing technique is based on the fact that solutions with mesh folds can still be useful for the exploration of the search space. MOREA considers mesh folds as being unrealistic and thus also considers solutions containing them to be invalid. Repairing these solutions is done by mutating each point in a folded tetrahedron and selecting the first mutation that resolves the fold, constrained by 64 attempts. The repair is only accepted if it is successful. Mesh folds will be discussed in further detail in Chapter 3. The adaptive steering technique in MOREA ensures the solutions are constrained to certain guidance objective values. MOREA therefore only preserves solutions between $[s_G{:}1.5s_G]$ after 100 generations, where s_G is the best guidance objective of any solution in the elitist archive.

2.5.4. Optimization Results

The output of the algorithm consists of the various deformed dual-dynamic meshes, each representing a solution along the approximation front. Additionally, each solution is represented by the three objective scores.

2.6. Deep Learning-based Landmark Detection

Deep Learning is based on the principle of using neural networks to learn representations from data (Haskins et al., 2020). In the context of DIR, this can be used to find corresponding landmarks, e.g. matching locations on both images that can be easily identified. See Figure 2.6 for an example. These landmark pairs can be used to guide a registration approach. Typically, landmark pairs are established manually by field experts. This results in large interobserver variation, however. Automatically finding landmark pairs to guide registration removes this variation.

The extraction of landmarks and the way they are used to perform DIR can be conducted in different ways. One way is to extract landmarks by segmenting a 3D CT image into body structures and using the information-rich environment of their boundaries to determine landmarks on the target image (Kearney et al., 2014). A small part of the area around these landmarks is used to identify similar volumes in the

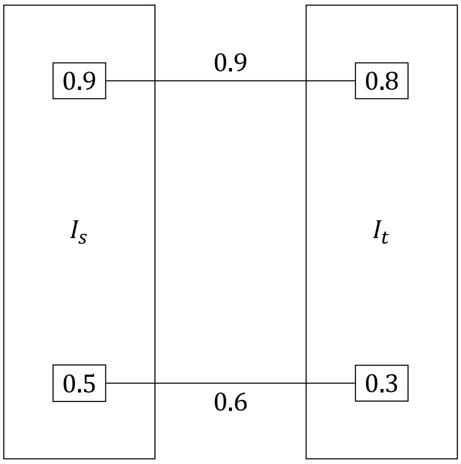


Figure 2.7: Example of a landmark pair set. Landmark probabilities are represented by node weights, while the descriptor probabilities are represented by edge weights.

source image, thus obtaining the predicted location of the landmark on I_t. After this, the landmark pairs are used as guidance in a Demons algorithm, to finally obtain the deformation map of the 3D CT image.

Another possibility is to extract landmarks based on features surrounding the landmark (Yang et al., 2017). Here, a 3D SIFT feature detector and 3D Harris-Laplacian corner detection are used. Landmarks are matched by first determining high-confidence matches based on a four-stage scheme where each stage represents a different resolution of the original image. After this is done, the high-confidence matches are used to guide the detection of new matches by leveraging their proximity to predicted landmarks within a certain radius. They mention, however, that the proposed method is not ready for automated DIR, as the landmarks still need to be manually verified.

A third approach is the Automated Landmark Detection approach (Grewal et al., 2020 and Grewal et al., 2023). Here, a convolutional neural network (CNN) is used to automatically detect landmarks that describe potentially relevant locations on the moving and fixed images. The approach then tries to find a rough mapping between landmarks on both images.

2.6.1. Preprocessing

Landmarks are generated using a Siamese neural network architecture containing CNN branches with shared weights, which is based on the U-net architecture (Ronneberger et al., 2015). Before the landmarks are generated, the input CT images, e.g., the moving and fixed image, are preprocessed by computing a “valid mask” for each image. This mask labels body voxels, to prevent processing on image regions outside the body. It is computed by applying an intensity threshold to the image and by removing small connected components in the binarized image.

2.6.2. Landmark extraction

The output of the network branches consists of a landmark probability map and feature descriptors of the image. The landmark probability map contains a prediction of possible landmark locations in image (I_s) and (I_t). A visual example of this can be viewed in Figure 2.7. Here, the landmarks on I_t have probabilities 0.9 and 0.5 of actually being a landmark, respectively. Their predicted pairs have probabilities of 0.8 and 0.3, respectively.

Both predictions are sent to a sampling layer, where landmark locations are sampled from the probability map according to the highest probability scores. From this, predicted landmark probabilities are extracted and feature-descriptor pairs generated, where for every $i \in \{1, 2, ..., K_1\}$ and $j \in \{1, 2, ..., K_2\}$, a feature descriptor pair $(f_i^{I_s}, f_j^{I_t})$ exists. Feature descriptor pairs are only generated for points where the corresponding valid mask point contains a 1.

2.6.3. Pair matching

The extracted feature-descriptor pairs are subsequently processed by a matching module, where the matching probabilities for each pair are predicted. A pair is considered matching during inference when its probability exceeds a predefined threshold. In the case that one landmark is matched to multiple pairs, the pair with the highest predicted descriptor probability is selected. In Figure 2.7, the descriptor probabilities are defined as the edge weights between the landmarks (in this case, 0.9 and 0.6, respectively).

2.6.4. Results

The approach returns a list of predicted corresponding source and target landmark pairs as physical coordinates. Additionally, each landmark pair contains their respective descriptor probabilities and the probability scores of the individual landmarks. These landmark pairs can then be passed on to a DIR approach for additional correspondence guidance.

3

Design

In this chapter, we give an overview of important components in the design of each hybrid. First, in Section 3.1 we describe the image pre-processing step for the clinical data used in this thesis. In Section 3.2, we discuss the use of landmarks as guidance information for the Digital Phantom. Finally, in Section 3.3, we explain how we use DVFs as a mesh initialization approach. As LGDP-MOREA contains each component, we have included a more detailed flowchart of this hybrid in Figure 3.1.

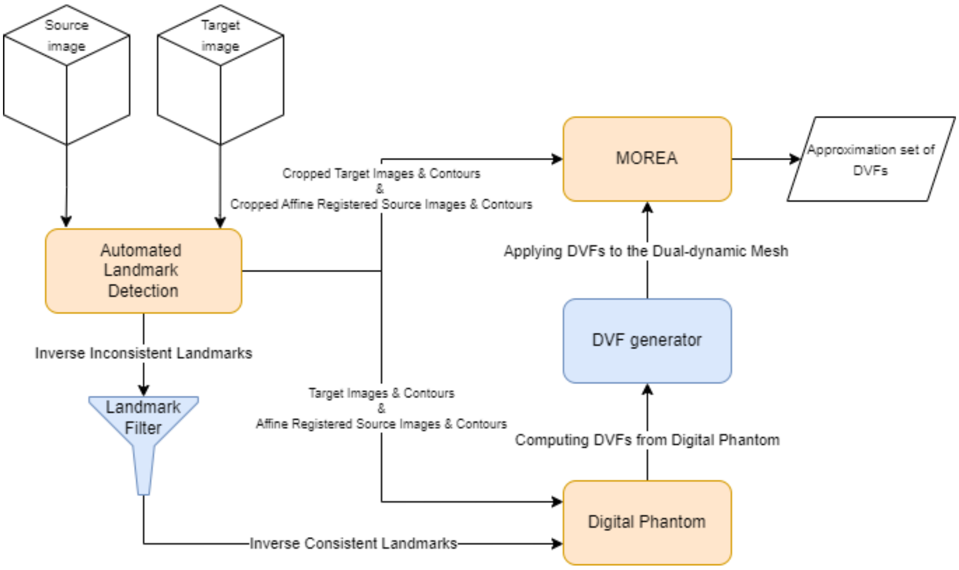


Figure 3.1: A flowchart describing the flow of the LGDP-MOREA pipeline. The original approaches are indicated in orange and the added direct modules are indicated in blue.

3.1. Image Processing

As can be viewed in Figure 3.1, the images that are used are processed in the Automated Landmark Detection pipeline. This includes an affine registration of the images and viewport cropping. Viewport cropping was performed only on the images that were passed to MOREA. The images used in the Automated Landmark Detection and Digital Phantom were not cropped. Viewport cropping was designed using the heuristic that we only need to use the bladder and its direct surroundings in optimization for this research. Having a smaller image to use reduces the computational load on MOREA and excludes irrelevant regions of the full scan.

To obtain the dimensions of the cropped viewport, we take the minimal and maximal annotation coordinates of the bladder and define a bounding box around it with a margin. We then crop the images and store them together with the dimensions and origin of the viewport to be used in the DVF generator. The images and their organ masks were resampled to a voxel spacing of 2x2x2mm before processing for the Automated Landmark Detection approach and MOREA. For the Digital Phantom, only the organ masks were saved, with a voxel spacing of 1x1x1mm.

3.2. Landmarks as Guidance Information

Landmarks on images I_s and I_t provide a point-to-point correspondence between the images, as described in Chapter 2. They can therefore provide information about the deformations between the two images. When evaluating an estimated transformation function T for this deformation, the distance between $T(I_s)$ and I_t can be used to measure similarity. In this research, we will denote these corresponding landmark pairs as $P(L_s, L_t)$ for I_s and I_t, where each landmark L has an x, y, and z coordinate.

This information could prove valuable as additional guidance information for the Digital Phantom approach. In a theoretical case, if a high-quality DVF could be found with high-quality landmarks, the transformation would transform the source landmarks closely onto the target landmarks. For DVFs of lower quality, there would be a mismatch between the transformed source landmarks and the target landmarks. This mismatch can be quantified using a distance metric. To be able to compute the distance between the transformed source landmarks and the target landmarks, the transformation of the landmark coordinates needs to be estimated with respect to the 3D tetrahedral models. The strategy used in this research is described in Section 3.2.2.

3.2.1. Inverse Consistent Landmarks

Preliminary experiments have shown that the landmarks resulting from the Automated Landmark Detection approach are not *inverse consistent*. We define inverse consistency of a set of landmark pairs $\mathcal{P}$ as:

$$\forall P(L_s, L_t) \in \mathcal{P} : \neg\exists P'(L'_s, L'_t) \in \mathcal{P} \mid (P'(L'_s, L'_t) \neq P(L_s, L_t) \wedge (L_s = L'_s \vee L_t = L'_t)) \tag{3.1}$$

This means that for any corresponding landmark pair P, there does not exist another corresponding pair P' that contains a source or target landmark from P. Figure 3.2 provides an example of a violation of inverse consistency, next to an example where inverse consistency is maintained. If the landmarks are not inverse consistent, one landmark might point to multiple locations, resulting in ambiguous guidance information. As several such cases were observed in preliminary experiments, we define a landmark filter to be applied before using the landmarks as guidance information.

To accomplish this, we leverage the fact that a pair P is not only defined by their respective landmarks but also by a score P_c representing the confidence that landmark L_s and L_t are corresponding landmarks. Keeping only the pairs for which the P_c is the maximum score for all pairs containing both L_s and L_t ensures the new $\mathcal{P}'$ meets Equation 3.1. This condition is defined as:

$$P(L_s, L_t) \in \mathcal{P}' \iff (\forall P'(L_s, L'_t) \in \mathcal{P} : P_c \geq P'_c) \wedge (\forall P'(L'_s, L_t) \in \mathcal{P} : P_c \geq P'_c) \tag{3.2}$$

In the case that there are multiple pairs with the same score, we use the summed probability scores of L_s and L_t as tiebreaker. The resulting landmarks each cover a point in 3D space. In the Digital Phantom approach, solutions are represented by 3D tetrahedral meshes. To be able to use the guidance information as a metric, the movement that each source landmark on the modeled source mesh makes when that modeled mesh is transformed needs to be described. This can be done using barycentric coordinate calculations.

3.2.2. Landmark transformation using barycentric coordinates

To estimate the transformation of the source landmarks, we calculate the barycentric coordinates of each landmark using the approach described in Chapter 2. To calculate the displacement of a source landmark using barycentric coordinates, it should lie within the volume of one of the tetrahedra in the source mesh. The relevant tetrahedron is identified by iteratively checking each tetrahedron in the mesh. During this check, subvolumes are formed between the landmark and the tetrahedron vertices. If the landmark lies within the tetrahedron's volume, these subvolumes do not exceed the total volume of the tetrahedron. This only occurs when the landmark L is outside of the tetrahedron. If a point lies

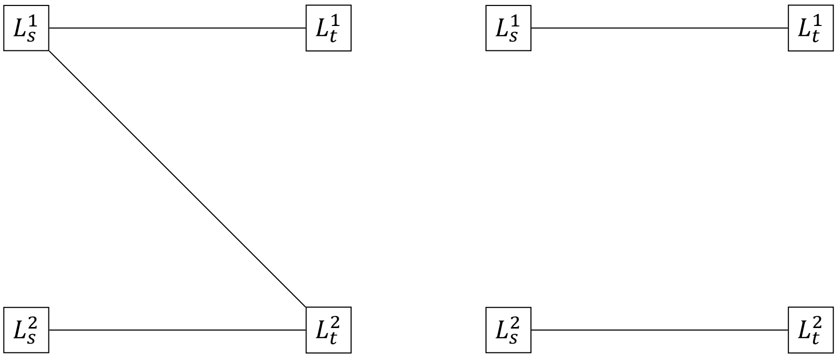


Figure 3.2: Example of an inverse inconsistent (left) and an inverse consistent (right) landmark pair set.

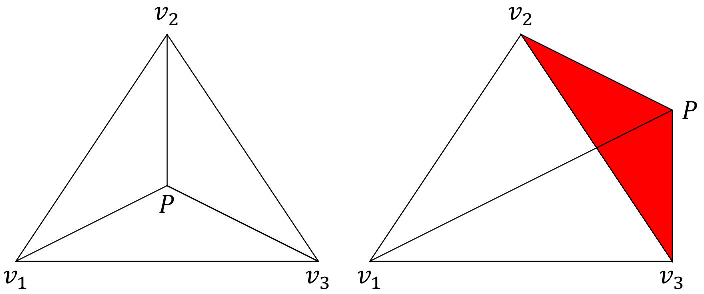


Figure 3.3: Example of a point P added to a triangle (left) and added outside of a triangle (right) with negative area highlighted in red.

outside a tetrahedron, the signed subvolume between the two points closest to the point becomes negative, according to Equation 2.4. Figure 3.3 contains a 2D example of such a case.

Using this information, the estimated transformation T on L can be applied by firstly using Equation 2.7 to compute ϕ_1, ϕ_2, ϕ_3, and ϕ_4. Since the vertices v_1, v_2, v_3, and v_4 of the tetrahedron are also moved, the new location of L is computed by solving:

$$T(L) = \phi_1 T(v_1) + \phi_2 T(v_2) + \phi_3 T(v_3) + \phi_4 T(v_4) \tag{3.3}$$

By iteratively solving this equation for each landmark in the source image, all landmarks are transformed to the target image based on the transformation of the tetrahedral mesh.

3.2.3. Distance Metric

After establishing the inverse consistent landmark pairs and calculating the transformed landmark positions, the guidance score of a transformation can be obtained by comparing the transformed source landmarks to the target landmark positions. For this, a distance metric needs to be defined. The Euclidean distance between two points p and q is a common metric for this purpose:

$$d(p,q) = \sqrt{(p_x - q_x)^2 + (p_y - q_y)^2 + (p_z - q_z)^2} \tag{3.4}$$

The complete guidance metric to be obtained can be expressed as the sum of all landmark distances. To properly be able to compare the obtained guidance metrics for different amounts of landmarks, we normalize the summed value, resulting in guidance G:

$$G(\mathcal{P},T) = \frac{1}{|\mathcal{P}|} \sum_{P(L_s,L_t)\in\mathcal{P}} (d(T(L_s),L_t)) \tag{3.5}$$

This guidance score can then be used as additional information in the Digital Phantom approach.

3.3. Using Deformation Vector Fields as Initialization

In MOREA, solutions are represented by a source and a target mesh. Each mesh has variable coordinates in the x, y, and z dimensions. In the initialization of the original implementation of MOREA, mesh points are strategically placed on organ contours, after which a random deviation is generated for each mesh point. Transforming the source or target mesh points using information obtained from the approximation set of the Digital Phantom, instead of applying random deviations, could provide a push in the right direction for MOREA, possibly improving convergence to high-quality registration solutions. To transfer the information necessary for this initialization, a DVF can be used, since this represents point-wise transformations in the voxel domain.

3.3.1. Computing DVFs from Digital Phantom

An approximation set of the Digital Phantom approach consists of different simulation parameter configurations, which in turn encode deformed 3D tetrahedral meshes. To obtain a DVF from a parameter configuration, we calculate the barycentric coordinates of every voxel in the target image.

MOREA uses a cropped version of the original source and target images to reduce computational complexity. Therefore, we only need to compute the DVF from these cropped dimensions. We define a bounding box from these dimensions where each point corresponds to a voxel from the MOREA field of view. These points have been transformed from voxel space to physical space, meaning that they overlap with the corresponding voxels from the three-dimensional mesh. Since the mesh is already in this physical space, the box represents points on the tetrahedra of the meshes. A visual representation of this bounding box can be viewed in Figure 3.4. From each of these points, we can use their barycentric coordinates together with the source vertex coordinates to calculate their transformed position from Equation 3.3. This information can be used to establish their deformation vectors.

Grouping landmarks by the tetrahedron in which they are located, for a large number of points, can be computationally expensive. If we compute the inverse DVF, i.e., the DVF describing transformation $T^{-1}(I_t)$, we need to perform this computation for every mesh in the approximation front. Since we want to minimize the computational complexity, we opted to calculate the forward DVF, i.e., the DVF describing transformation $T(I_s)$. This allows us to only compute this grouping and the barycentric coordinates once, on the mesh of I_s. The transformation of the landmarks can be found per mesh in the approximation set by using Equation 3.3.

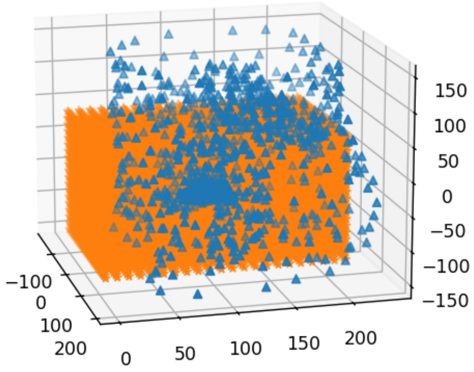


Figure 3.4: An example of the bounding box used to define a DVF from a tetrahedral mesh. The bounding box points are represented by orange stars, while the tetrahedral mesh points are represented by blue triangles.

3.3.2. Applying DVFs to the dual-dynamic mesh

Each solution in MOREA is represented by two meshes; the source and target mesh. Deformation is modeled by the difference in location between corresponding mesh points. Intuitively, for a true deformation, we want the corresponding tetrahedra formed by the points to contain the same structures in I_s as in I_t. An illustration of this is given in Figure 3.5. Currently, the meshes are initialized with respect to the organ contours of the source image. This mesh is copied to the target image. To ensure each tetrahedron of the target mesh captures as much of the same information on the corresponding tetrahedron of the source image as possible, we apply the obtained forward DVFs to the target mesh points.

DVF selection

MOREA typically has a population of 700 solutions. Since the set of DVFs resulting from the Digital Phantom can be smaller or larger than the target initialization population size of MOREA, solutions are generated as follows: First, we select a maximum of 250 DVFs from the DVF set. This is done by sorting the DVFs by increasing Dice score, and skipping n DVF solutions until 250 DVFs are found. n is determined by dividing the total DVF set size by 250 and taking the nearest lower integer. This approach ensures we are still exploring the entire approximation set of DVF solutions in our initialization. In case the number of DVFs is less than 250, we select the entire set.

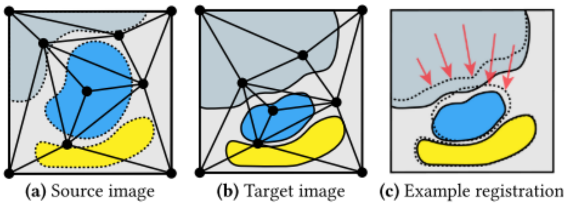
DVF Application

When the forward DVF set is selected, it is applied to the target meshes without additional deviation of the source and target meshes. This ensures the deformation contained in the DVFs does not accidentally get modified due to random deviation, which could undo the gain in information from applying the DVFs. After the full set of DVFs has been applied, it is re-applied as many times as necessary to reach the population size of MOREA. At every re-application, a multiplication factor is used to scale the DVFs. This multiplication factor is determined by Equation 3.6. Here, n represents how many re-applications have taken place before the current re-application. Additionally, a random deviation in the source and target meshes is applied for these re-applications. This random deviation is applied using the Gaussian kernel initialization approach from MOREA with a magnitude factor. This magnitude is inversely related to the multiplication factor. This finally leads to a set of 700 solutions which have each been initialized with a DVF, but which are also diverse due to the random deviation strategy.

$$multiplication\_factor = 1 - \frac{|DVFs|}{|population|} * n \tag{3.6}$$

Preventing constraint violations in DVF initialization

Preliminary results of the Digital Phantom showed inverted volumes or mesh folds present in the three-dimensional mesh of the solutions in the approximation front. Calculating a forward DVF from these meshes and applying them to the target mesh of MOREA could therefore also introduce mesh folds. MOREA considers these mesh folds to be constraint violations and filters solutions containing two or more folds out of the population. To prevent mesh folds from occurring in the initial population due to our application of DVFs, we leverage the step-wise initialization method of MOREA. Instead of applying a deformation vector at any point with its full magnitude, we divide this application into a number of steps, where every step a fraction of the magnitude of the DVF vectors is applied to the mesh. At every step, we check whether a constraint violation occurs for each point that gets moved and do not move this point in case it does, for this step. This results in a target mesh that generally follows the deformation of a given DVF, but does not contain any constraint violations.



(a) Source image **(b)** Target image **(c)** Example registration

Figure 3.5: A two-dimensional example of the source mesh (left) and target mesh (middle) having roughly the same bladder contents per triangle. (Andreadis et al., 2023)

4

Experimental Setup

In this section, we will explain the methodology behind the evaluation of the individual approaches and their hybrids. We include both a quantitative and qualitative evaluation of registrations. In this section, we first discuss how the quantitative evaluation was performed. Next, we examine the methodology of the qualitative evaluation. Finally, we describe the specifications of the performed experiments.

4.1. Dataset

Evaluation was performed on pelvic CT scans of three cervical cancer patients, acquired at the Leiden University Medical Center (LUMC). For each patient, a scan with an empty bladder and a scan with a full bladder was included, acquired originally for treatment planning purposes. These scans therefore feature large deformations in the pelvic region, largely caused by the change in bladder filling. In our work, we only explicitly model the bladder, the bones, and the body of the patient. The bladder was chosen over other organs since it can have the largest volumetric change between the source and target image. In every evaluation, we modeled the deformation from a low energy state to a high energy state, meaning that the volume of the source bladder is less than the volume of the target bladder for every patient. The scan pairs are pre-processed as described in Chapter 3.

4.2. Configurations

We have split up the evaluation of the Landmark Guided Digital Phantom into three separate configurations, to test the introduced filter strategy and how the landmarks are integrated in the Digital Phantom. First, we consider replacing the deformation energy objective of the Digital Phantom with the filtered guidance metric. Then, we consider the same setting but without filtering the landmarks for consistency. Finally, we consider replacing the Dice score of the Digital Phantom with the filtered guidance metric instead. We call these configurations “Dice and Landmarks with Filter”, “Dice and Landmarks without Filter”, and “Deformation Energy and Landmarks with Filter”, respectively.

Preliminary experiments showed the last sub-evaluation was highly outperformed by the first two sub-evaluations. We have therefore opted to later only evaluate the full hybrid, LGDP-MOREA, on the Dice and Landmarks with Filter configuration. We only use the filtered landmarks, as the unfiltered landmarks tend to contain inconsistent landmark matches.

4.3. Quantitative evaluation

The output of each component, and therefore also of each hybrid, has a different form for every approach. The Automated Landmark Detection approach returns a set of corresponding landmarks, and the Digital Phantom and MOREA both return a set of deformed 3D meshes. Additionally, the metrics upon which each output is based are different, which makes a quantitative comparison between components difficult. We therefore derive DVFs from all approaches and re-evaluate these DVFs with a set of independent metrics.

4.3.1. DVF Estimation

To be able to do a proper comparison between different DVFs, we need to ensure the spacing and size of the DVFs are all the same. MOREA uses a cropped version of the CT scans used in the Automated Landmark Detection and Digital Phantom approaches, so we only estimate the DVF within the viewport of this cropped version, meaning that we use the same image origin and dimensions. We ensure the DVFs all have spacing 2x2x2mm. In the case of this evaluation, every generated DVF is an inverse DVF, representing the voxel movements from a CT scan containing a full bladder towards one with an empty bladder.

The inverse DVFs from the Automated Landmark Detection approach were derived using Elastix, where the B-spline optimization method was used to estimate the deformations based on the generated landmarks. In this B-spline-based DIR method, we used the `CorrespondingEuclideanDistanceFilter` metric with weight 1 and the `TransformBendingEnergyPenalty` metric with weight 70. The first metric tries to match the landmarks as closely as possible, while the second regularizes the deformation. Weighting these metrics was done exploratively by looking at individual solutions with different regularization weights.

To generate the inverse DVFs from the front of the Digital Phantom, we apply the same strategy as described in Section 3.3. The DVFs from MOREA are generated based on the relative mesh point movements between the source and target meshes.

The metrics upon which we base the quantitative analysis also have to be equal for all components. Therefore, we compute two new metrics from the calculated inverse DVFs. One of these metrics is based on Hooke's Law and the other is called the Hausdorff Distance. We describe both, below.

4.3.2. Hooke's Law

The metric based on Hooke's Law is a measure of deformation energy. For every neighboring vector pair, the magnitude of the differences is calculated. We use this metric since we are interested in how much deformation occurs within one solution and how homogeneous the DVF is. We consider a more homogeneous DVF as a more stable deformation, as there are fewer sudden displacements between neighboring voxel vectors. This energy can be defined using Equation 4.1 (Andreadis et al., 2024):

$$f_{def.magnitude} = \sum_{p_t \in P_t} \left[\frac{1}{|n(p_t)|} \sum_{a \in n(p_t)} ||T'_{p_t} - T'_a||^2 \right] \tag{4.1}$$

Here, T'_{p_t} is defined as the deformation vector at point p_t in the DVF, and we use $n(p_t)$ to denote all directly adjacent neighbors of point p_t. In this definition, the adjacent neighbors are all voxels that are directly neighboring p_t, excluding diagonal neighbors. Since the naive implementation of this calculation is computationally expensive for the large DVF images that we use, we compute the metric using matrix-wise operations.

4.3.3. Hausdorff Distance

We use the Hausdorff Distance as a measure of contour similarity. The metric generally represents the worst minimal distance between any point in set A and its closest point in set B and vice versa. This is defined as follows (Huttenlocher et al., 1993):

$$H(A, B) = \max(h(A, B), h(B, A)) \tag{4.2}$$

and

$$h(A, B) = \max_{a \in A} \min_{b \in B} ||a - b|| \tag{4.3}$$

A visual example of this definition can be viewed in Figure 4.1. In this example, the distance metric, denoted by $||a - b||$, is the Euclidean Distance between point a and b. To compute the Hausdorff Distance score of an inverse DVF, we use it to transform the target bladder contour annotations. We then compare the transformed target bladder annotations with the source bladder annotations using Equation 4.2. Since the Hausdorff Distance represents the worst distance between two sets, it is prone to be influenced by outliers. To remedy this, we opted to use the 95th-percentile Hausdorff Distance.

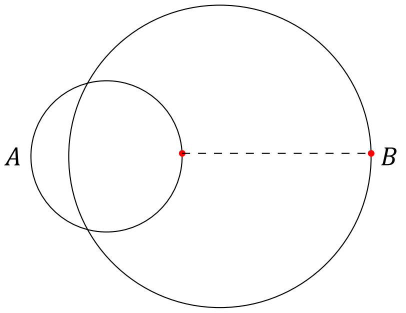


Figure 4.1: Example of the points selected for the Hausdorff Distance between a set A and a set B. The Hausdorff Distance corresponds to the length of the dashed line.

4.3.4. Quantitative Evaluation Settings

Since each instance of a run is based on a random seed, we ensure a fair comparison by conducting 5 repeat tests per patient. We then select one reference point for all repeat tests and compute the hypervolume metric with respect to that reference point (Guerreiro et al., 2021). The hypervolume acts as an indicator of the quality of an approximation front. A larger hypervolume indicates a more diverse, higher-quality front. To compare MOREA with DP-MOREA and LGDP-MOREA, we conduct a one-sided Mann-Whitney U test to test for statistical relevance. We do this per patient, as we cannot statistically compare the results of different patients with one another. We also apply Bonferroni correction, reducing the α-value with a factor 2 for statistical significance, from 0.05 to 0.025.

4.4. Qualitative Evaluation

We evaluate the solutions qualitatively by looking at how well the contours of the images overlap, and whether any big unexpected changes in the image occur. We select the solutions to evaluate by exploring the approximation front starting at the minimal guidance solution while looking at the criteria established above. In a clinical setting, this would also be done this way, as the user of a Multi-Objective EA has to do this *a posteriori* themselves when selecting a solution. As one image also contains many different slices, we will select the ones that visually represent the target best.

4.5. Specifications

4.5.1. Automated Landmark Detection

The Automated Landmark Detection approach was run using a landmark confidence threshold and descriptor confidence threshold of 0.5 each. The input images were not cropped and were read with a voxel spacing of 2x2x2mm.

Machine Specifications

Every repeat had access to one NVIDIA RTX A5000 or RTX A6000 Graphics Processing Unit (GPU) with 8 GB of allocated Random Access Memory (RAM).

4.5.2. Digital Phantom

Version

The version of the Digital Phantom used in this evaluation differs from the one described by the original publication (Rodriguez et al., 2023). Instead of using MAMaLGaM, the code obtained from the author uses RVEA to optimize the simulation parameters. Additionally, the simulation parameters were limited to the internal pressure of the bladder, the external orthogonal forces on the bladder, the orthogonal forces acting on the body and the tissue elasticity. This results in a total of eight optimizable parameters instead of the original twelve.

Mesh Generation

To generate the meshes used in the Digital Phantom, we first convert the voxel-based body, bladder, and bones to surface triangular meshes. A marching cubes algorithm is used to determine how fine-grained the surface meshes are. We set the marching steps to 4 for the body mesh and 1 for the bladder mesh. After this step, we use the `fast-simplification`[1] package to simplify the meshes with a weight of 0.99.

When the surface meshes are established, we create a list of fixed nodes representing the bony anatomy of the patient by subsampling 500 nodes from the surface mesh of the bones. After this, the TetGen Suite[2] is used to generate a tetrahedral mesh using a Delaunay tetrahedralization. The triangular surface meshes of the body and bladder and the selected bone nodes are used as constraints in this process. To balance mesh quality with mesh node additions, we set the Steiner budget to 50.

Optimization Loop Settings

RVEA was run with a population size of 16 with an evaluation budget of 10,000. For every repeat, we use the same generated mesh. The random seeds for each repeat were applied during the optimization loop of RVEA.

Machine Specifications

The Digital Phantom was also run on a cluster, on which most nodes run on an Intel E5-2697 or Intel E5-2690. Each repeat had access to 16 Central Processing Units (CPUs) and 8GB of memory per CPU.

4.5.3. MOREA

Version

The version of MOREA used in this evaluation also differs from the one described by the original publication (Andreadis et al., 2023). In this version, the custom tissue elasticities are not used. Instead, Equation 2.2 is adapted to:

$$f_{magnitude} = \frac{1}{10|\Delta|} \sum_{\delta \in \Delta} \left[\sum_{(e_s, e_t) \in E_\delta} (||e_s|| - ||e_t||)^2 \right] \tag{4.4}$$

Additionally, MOREA's contour-based initialization was adapted to only generate meshes based on the bladder and bone contours of each patient, instead of on the whole body.

Parameter Settings

MOREA was run using a generation budget of 300, a population size of 700 and 300 mesh points. Any other parameters were taken from the parameter configuration recommendation of MOREA (Andreadis et al., 2023).

Machine Specifications

Each repeat of MOREA was run on one NVIDIA RTX A5000 or RTX A6000 GPU, with 30GB of allocated RAM.

---

[1] https://pypi.org/project/fast-simplification/

[2] https://pypi.org/project/tetgen/

5

Results

In this chapter, we discuss the results of the experiments described in Chapter 4. We first consider the individual Automated Landmark Detection approach, followed by a comparison of the Digital Phantom and the proposed LGDP configurations. After this, we compare the results of MOREA with those of DP-MOREA and LGDP-MOREA. Finally, we provide a full comparison of the individual approaches.

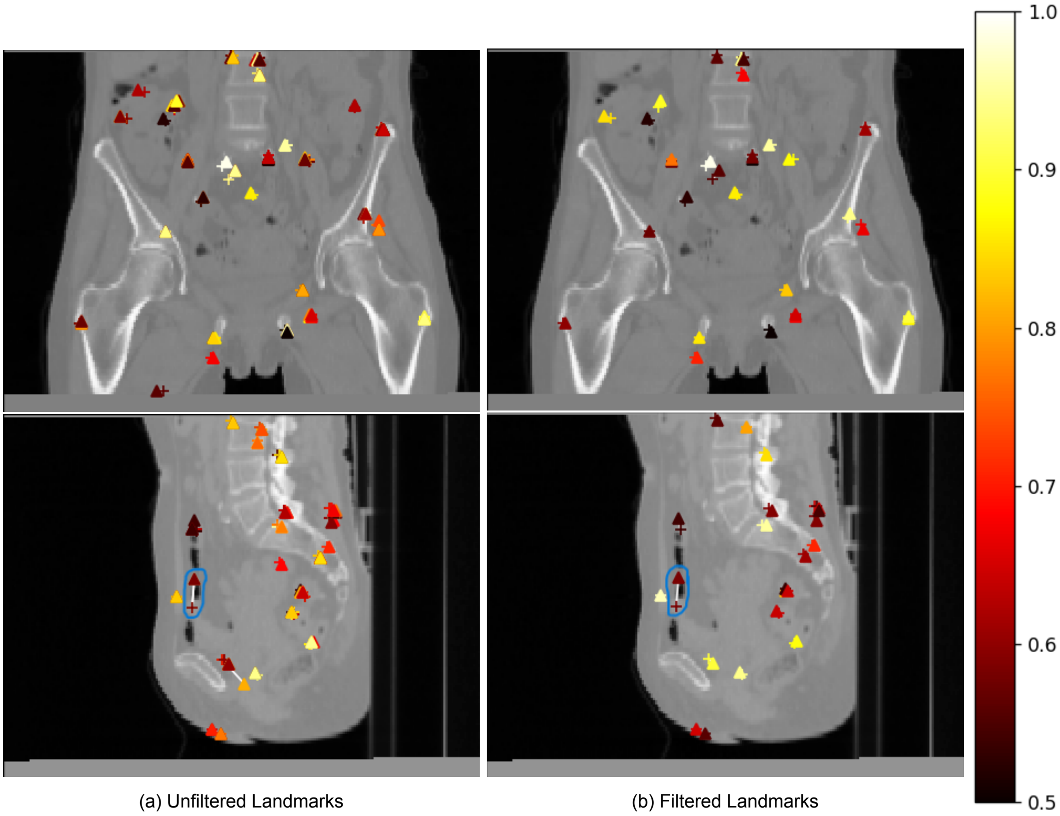
5.1. Automated Landmark Detection

In Figure 5.1, we display a slice of both the filtered and unfiltered landmarks from the Automated Landmark Detection approach. Landmarks are most useful and of high quality if there are many landmark pairs directly around the source bladder describing a displacement toward the target bladder. This provides more information about the desired displacements of the bladder, which we are optimizing as mentioned in Chapter 4. Additionally, landmark pairs with both high source and target confidence scores are expected to be more accurate and therefore more informative. Finally, high-quality landmarks are expected to be inverse consistent, to avoid inconsistent displacement information.

We observe that the landmark pairs mostly describe small displacements, both for the unfiltered and filtered landmarks. This is also visible in Figure 5.2. There also seem to be few landmarks near the bladder contour. These landmark pairs also do not always describe the actual displacement from the source bladder to the target bladder.

Some predicted deformations are likely to be unrealistic, for example near the gas pockets in the bowel, indicated by the blue circle in the sagittal view. These are visible for both the unfiltered and filtered landmarks, indicating that these are not filtered out. We also observe there is a slight visual difference between the confidence scores of the landmarks in the unfiltered and filtered approach. This visual difference is most likely due to the scaling of the displayed images, because of which some inverse inconsistent landmark pairs describing small different displacements are displayed on top of one another. The visual difference between the unfiltered and filtered landmarks indicates that these inverse inconsistent landmark pairs are being filtered out of the landmark set. This shows that the landmark filter indeed has an observable effect on the landmarks. There is no clear evidence that filtering the landmarks increases or decreases the average confidence of the remaining pairs, however.

We display the deformed target bladder contours using the DVFs generated with Elastix from the generated landmarks in Figure 5.3. Since the DVFs are inverse, the target bladder contours are transformed and compared with the source bladder contours. In a case with ideal landmarks, we expect the deformed target bladder contours to overlap well with the source bladder. For Patient 2, the generated deformation with the filtered landmarks seems to create a better-fitting deformed bladder contour than that of the non-filtered landmarks. For Patient 3, however, there is no clear better-fitting bladder contour. For both unfiltered and filtered landmarks, the deformed target bladder does not overlap well with the source bladder, however. This indicates that both filtered and unfiltered landmarks in their current form are not well-suited to be directly used in combination with the Elastix DIR method. Furthermore, we observe that using the landmark filter can influence the registration result, as is visible for Patient 2. Here, the deformed bladder contour fits the source bladder contour better when filtered landmarks are used than unfiltered landmarks.



(a) Unfiltered Landmarks (b) Filtered Landmarks

Figure 5.1: Landmarks from the Automated Landmark Detection approach in the coronal (top) and sagittal (bottom) source view of Patient 1. Source landmarks are displayed with plus signs, target landmarks are represented by triangles. Landmark pairs are connected by a line. One seemingly unrealistic displacement is highlighted by a blue circle. A landmark's color represents the confidence that it is a landmark.

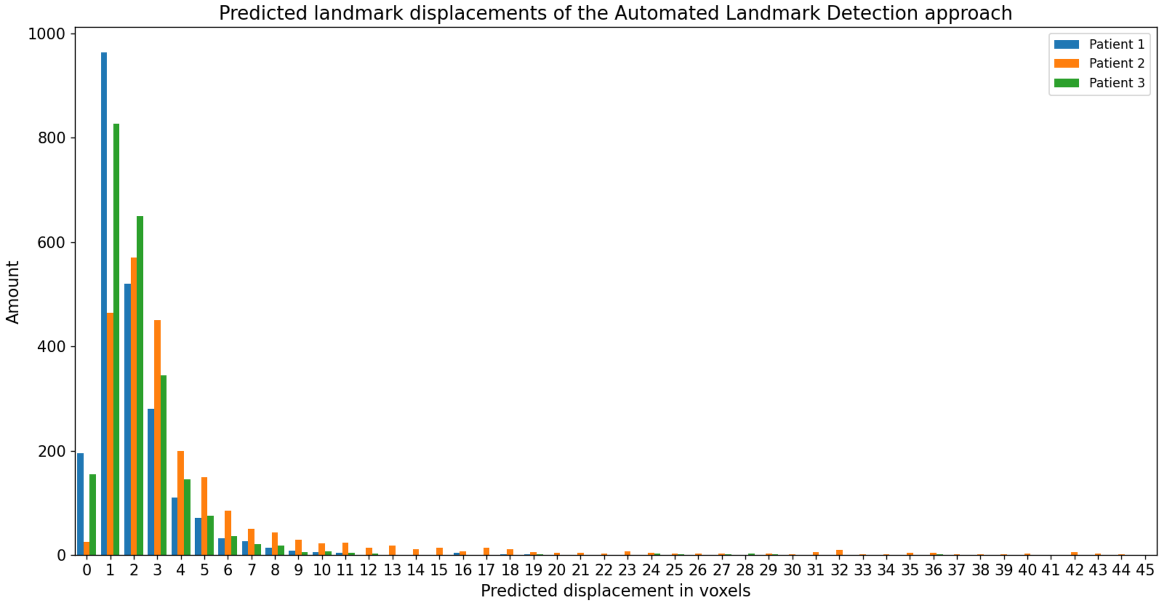
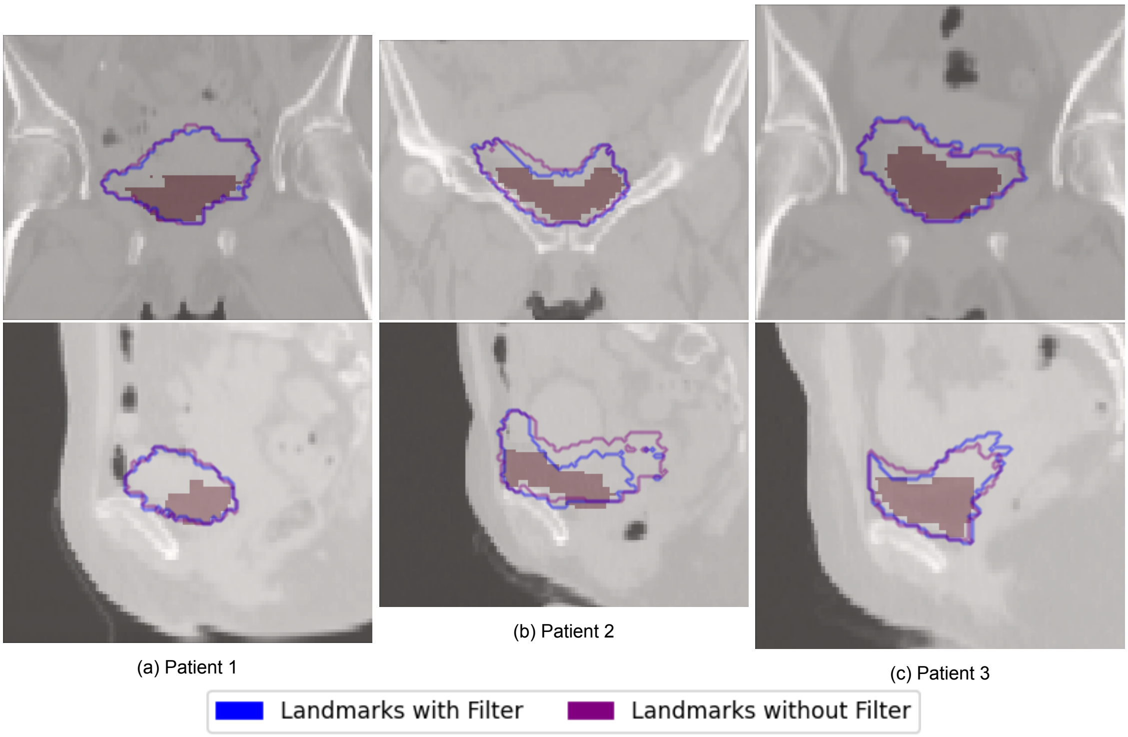


Figure 5.2: Amount of filtered landmarks with a predicted displacement for the Automated Landmark Detection Approach. The voxel spacing is 2x2x2mm.



(a) Patient 1

(b) Patient 2

(c) Patient 3

Figure 5.3: Deformed bladder contours using the landmarks from the Automated Landmark Detection with and without filter on the source image for all patients with the source bladder in full red in coronal (top) and sagittal (bottom) view.

5.2. (Landmark Guided) Digital Phantom

5.2.1. Convergence of Optimization Objectives

In Figure 5.4, and in Figures A.1 and A.2 in the Appendix, the convergence curves with respect to the optimization objectives are shown. In these figures, we have included all repeats. These plots show the approximation front after different generations, giving an indication of how many generations

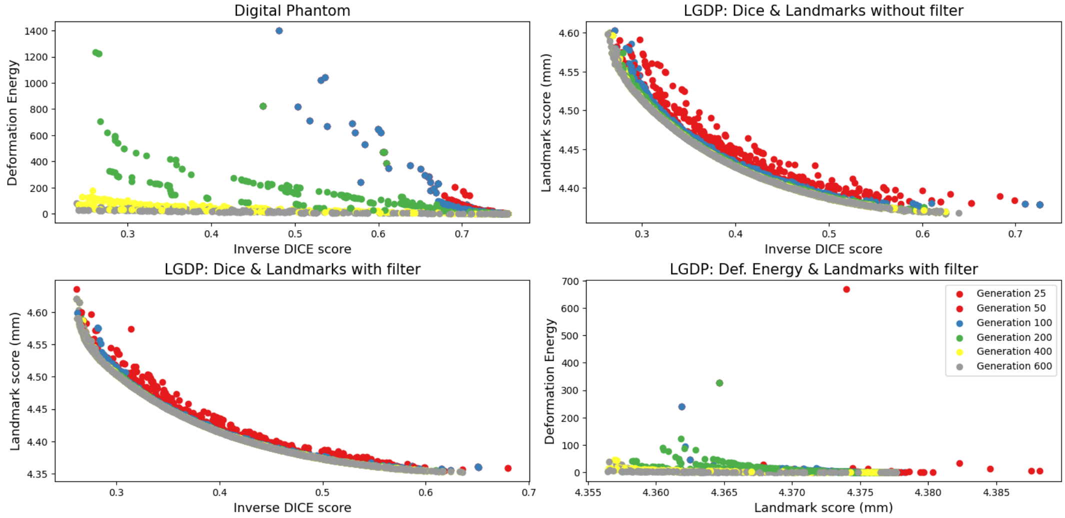


Figure 5.4: Convergence in terms of the original objectives of the Digital Phantom for Patient 1.

it would take to converge to the final approximated front. We expect to see the landmark scores and deformation energy to form an effective tradeoff, due to the guiding nature of the landmark scores.

An interesting observation on these results is the shape of the curves when using the landmarks as guidance objectives instead of the Deformation Energy. For each patient, it clearly shows a tradeoff between the Dice and Landmark objectives. For the LGDP Deformation Energy with Landmarks configuration, we observe that barely any deformation occurs, even though a tradeoff is slightly visible. In the meantime, the landmark scores obtained by this configuration when compared with the LGDP Dice and Filtered Landmarks configuration are lower for each patient. This indicates that the deformed source landmarks match better with the target landmarks for the LGDP Deformation Energy with Landmarks configuration, while still barely deforming the mesh. The landmarks mostly describe very small displacements, as indicated in Figure 5.2, however. This indicates that the landmarks, in their current form and implementation, would act more as a regularizing objective instead of guidance information.

When looking at the number of generations required to reach convergence, we notice that both LGDP Dice with Landmark configurations move faster toward their found approximation front than the other configurations. The Baseline Digital Phantom and the LGDP Deformation Energy with Landmarks configuration seem to have a similar convergence speed. When comparing the LGDP Dice with Landmark configurations, we observe that the filtered landmarks have less variance in between repeats during convergence in Figure A.2.

5.2.2. Hooke’s Law and Hausdorff Distance scores

In this subsection, we quantitatively evaluate the solutions using the Hooke’s Law and Hausdorff Distance scores as independent evaluation metrics. In Figure 5.5, the values for the Hooke’s Law and Hausdorff Distance metrics have been plotted for each DVF from the approximation set of the Digital Phantom and the different LGDP hybrid configurations. Each individual line represents a repeat. Again, we expect to see a tradeoff between the Hooke’s Law and Hausdorff Distance metrics, as these are forms of the regularizing deformation magnitude and the matching contour similarity metrics. This tradeoff is also represented by a convex curve.

We observe that the Digital Phantom baseline configuration is able to achieve lower Hausdorff Distances while maintaining a lower Hooke’s Law score compared to all other configurations for all patients. The LGDP Deformation Energy with Landmarks configuration is dominated by all other configurations, as it highly underperforms for both metrics. This reinforces the hypothesis that the landmarks act more as a regularization objective than a guidance objective in their current form. Both LGDP Dice with Landmarks configurations seem to have comparable performances. They seem to follow the same trend as the baseline Digital Phantom, except for Patient 3. A more thorough examination of the solutions of this patient revealed the existence of large inverted volumes in the patient mesh, which cause larger Hausdorff Distances. Figure 5.6 shows an example of this.

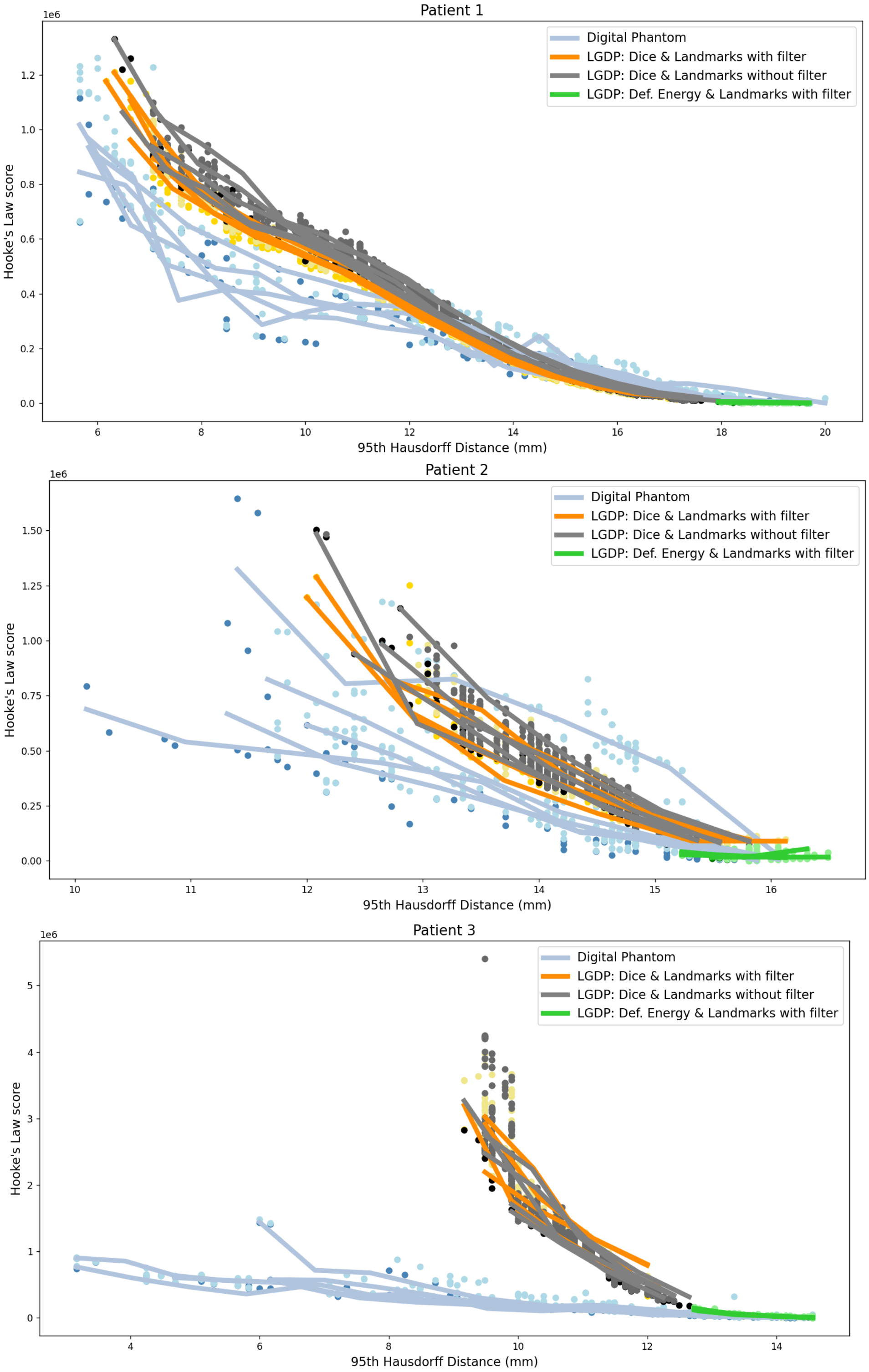


Figure 5.5: Calculated Hooke and Hausdorff values for the DVF approximation set of the Digital Phantom. Darker colors indicate non-dominated solutions in each repeat.

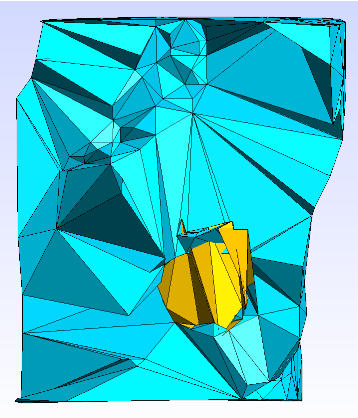


Figure 5.6: Deformed tetrahedral mesh representation of Patient 3, from a side perspective (sagittal). Bladder tetrahedra are indicated in yellow.

Based on the Hooke's Law and Hausdorff scores, we calculate the hypervolume for every repeat with the same reference point ($[25, 10^7]$). The resulting box plots can be viewed in Figure 5.7. A configuration with better performance is indicated by a higher hypervolume. We observe that in every tested case, the baseline Digital Phantom outperforms all other configurations in terms of the calculated hypervolume. Furthermore, the LGDP Deformation Energy with Landmarks configuration is dominated by all other configurations. We observe that the LGDP Dice with Landmarks configurations both have comparable performances quantitatively.

5.2.3. Qualitative Evaluation

In Figures 5.8, 5.9, and 5.10, the selected inverse DVFs of each configuration are visualized together with the deformed target bladder. We observe that the deformation magnitude for the LGDP Deformation Energy with Landmarks configuration for each patient is very low compared to the other configurations. When we combine this observation with the deformed target bladder contours in Figure 5.11, we can conclude that the deformation is minimal. This does not hold for the other LGDP configurations.

For both the Digital Phantom and the LGDP configurations, deformation mainly occurs in the area of the bladder contours. The DVFs of the Digital Phantom and the LGDP Dice with Landmarks configurations mainly differ in their homogeneity. The LGDP Dice with Landmarks configurations both contain a sudden large deformation in voxels directly around the bladder, however. Considering the deformed target contours from the Digital Phantom and the LGDP Dice with Landmarks configurations, we observe that the obtained DVF and contours are more homogeneous in their shape and magnitude for the Digital Phantom. The LGDP Dice with Landmarks configurations contain some large outliers in the contours, while the contours of the baseline Digital Phantom remain relatively smooth, especially for Patient 1 and 3. This indicates that the baseline Digital Phantom can visually be considered more stable. This can be due to the fact that LGDP introduces large inverted volumes in the deformed mesh. When we create a DVF based on this mesh, the inverted volumes can describe unrealistic displacements.

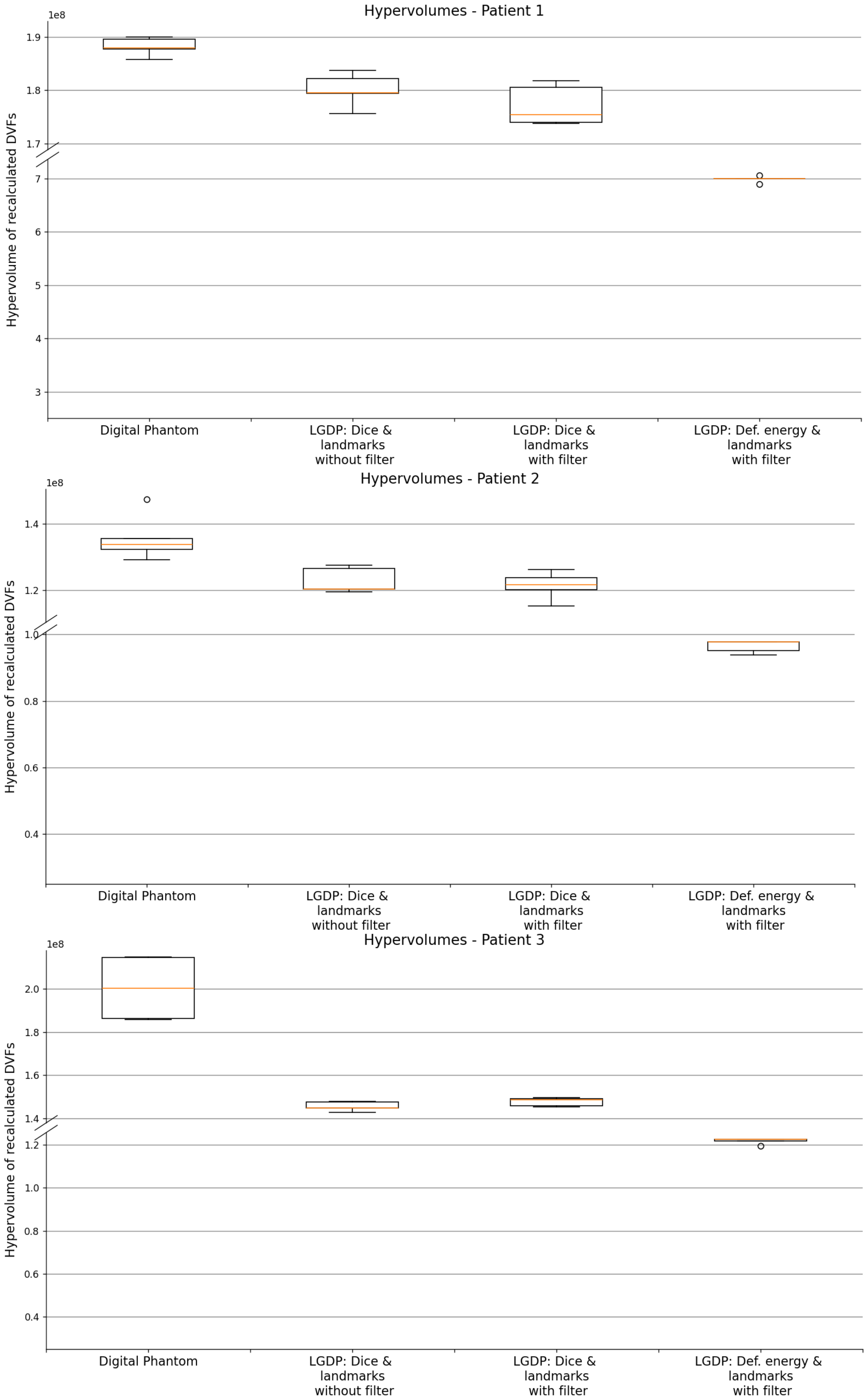
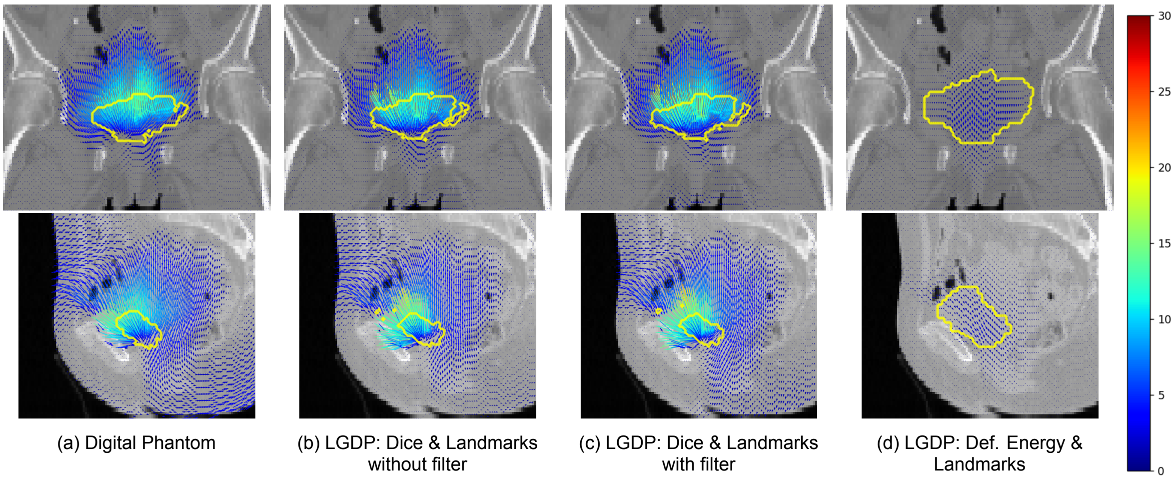


Figure 5.7: Box plots of Digital Phantom hypervolumes, calculated from the Hooke's Law and Hausdorff scores. Each individual hypervolume represents one repeat. A break has been inserted in the vertical axis.



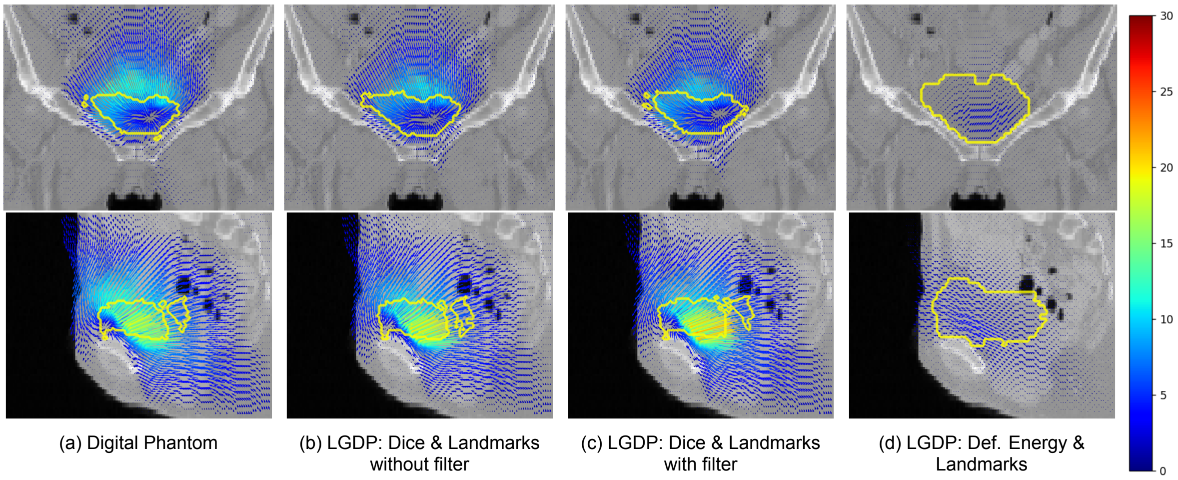
(a) Digital Phantom

(b) LGDP: Dice & Landmarks without filter

(c) LGDP: Dice & Landmarks with filter

(d) LGDP: Def. Energy & Landmarks

Figure 5.8: Downsampled DVFs and the deformed bladder contours of the Digital Phantom configurations on Patient 1 in coronal (top) and sagittal (bottom) view. The arrow colors represent the vector magnitudes in voxel units.



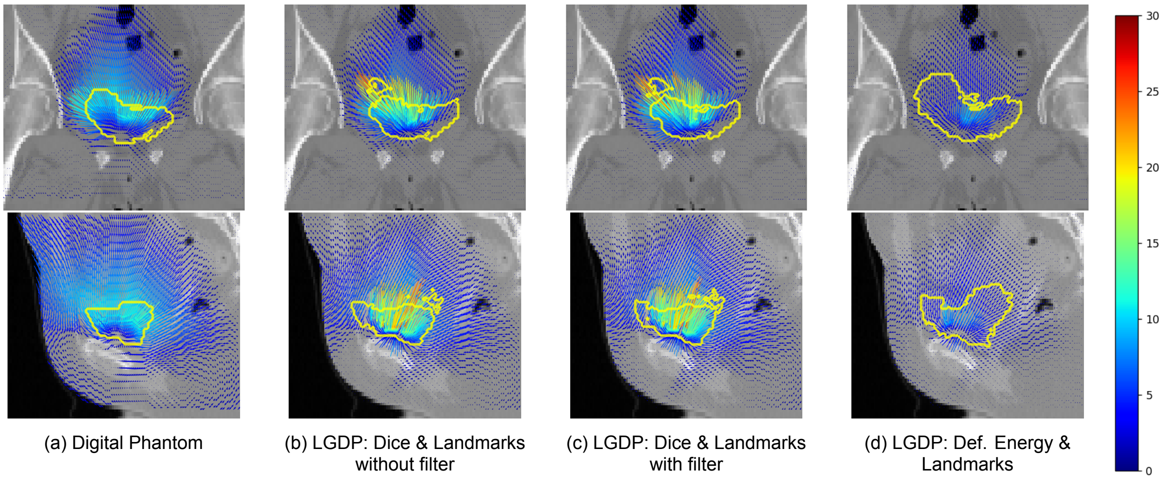
(a) Digital Phantom

(b) LGDP: Dice & Landmarks without filter

(c) LGDP: Dice & Landmarks with filter

(d) LGDP: Def. Energy & Landmarks

Figure 5.9: Downsampled DVFs and the deformed bladder contours of the Digital Phantom configurations on Patient 2 in coronal (top) and sagittal (bottom) view. The arrow colors represent the vector magnitudes in voxel units.



(a) Digital Phantom

(b) LGDP: Dice & Landmarks without filter

(c) LGDP: Dice & Landmarks with filter

(d) LGDP: Def. Energy & Landmarks

Figure 5.10: Downsampled DVFs and the deformed bladder contours of the Digital Phantom configurations on Patient 3 in coronal (top) and sagittal (bottom) view. The arrow colors represent the vector magnitudes in voxel units.

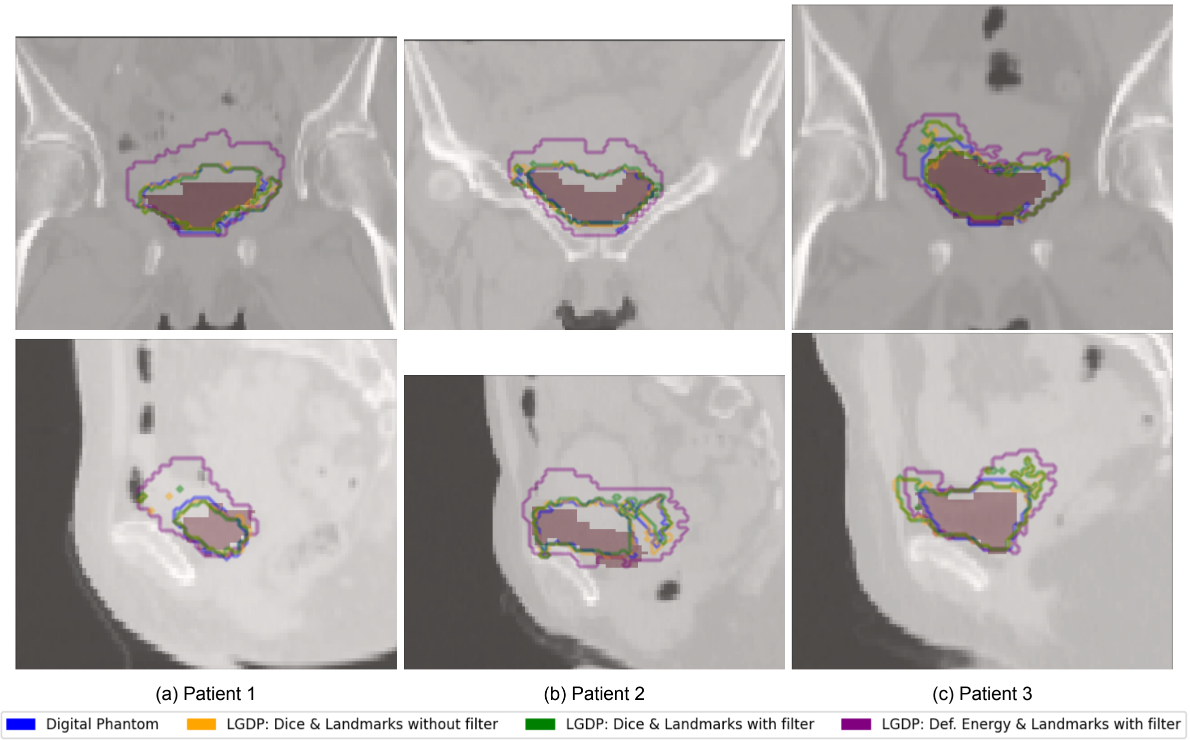

(a) Patient 1
(b) Patient 2
(c) Patient 3

Figure 5.11: Deformed bladder contours on the source image for all patients and Digital Phantom variants with the reference bladder in full red in coronal (top) and sagittal (bottom) view.

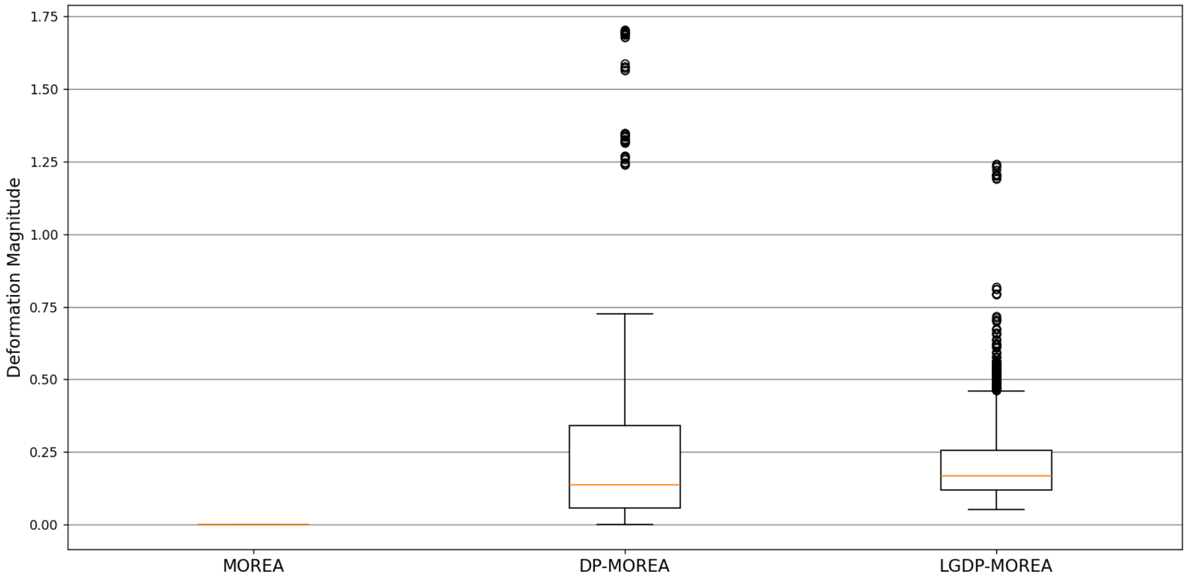
5.3. (LGDP)-MOREA

5.3.1. Initialization

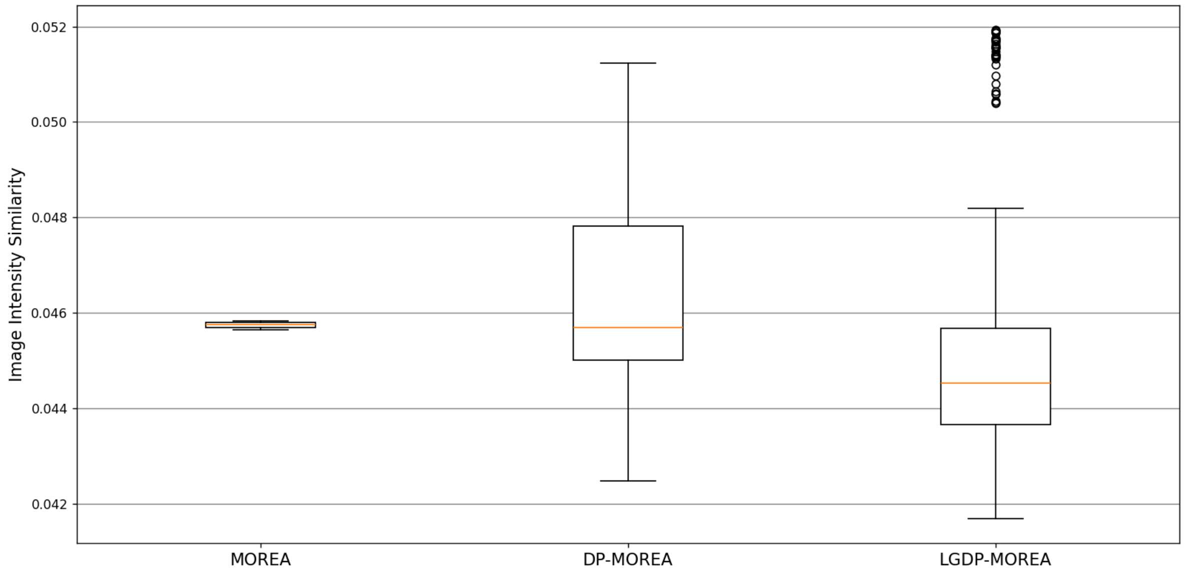
We first analyze and compare the initialization of MOREA with DP-MOREA and LGDP-MOREA. We have selected one repeat and one patient to investigate the objective values in the initial population. These are displayed in Figure 5.12. If the initialization using the DVFs from the Digital Phantom indeed has an effect on the population, we expect to see generally larger deformation magnitudes as computed by MOREA in DP-MOREA and LGDP-MOREA. Since we apply a multiplication factor in the initialization, we also expect the variance in this objective to be wide for the population. Additionally, we expect a generally lower guidance score for DP-MOREA and LGDP-MOREA when the first population has been initialized. Since we are minimizing the guidance objective, a lower guidance score is considered better than a higher guidance score.

We observe that the deformation magnitudes for the initialized populations of DP-MOREA and LGDP-MOREA indeed are larger than that of MOREA. DP-MOREA seems to sometimes contain individual solutions that have a very large deformation magnitude. When considering the image similarity scores, we also see a wide variance in solutions in both the positive and negative directions for DP-MOREA and LGDP-MOREA. Interestingly, this initialization does not result in an image intensity similarity that is always lower than MOREA. This could be due to the fact that optimization with the Digital Phantom or LGDP is not based on image similarity and is mostly focused around the bladder. Initializing a population with a DVF based on the Digital Phantom or LGDP will thus not necessarily result in a better image intensity similarity score. In terms of the guidance scores, DP-MOREA and LGDP-MOREA indeed both seem to reach lower values in the initialization than MOREA. This indicates that the informed initialization using the Digital Phantom indeed does have a positive effect on both variation in the initial population and the initial guidance objective.

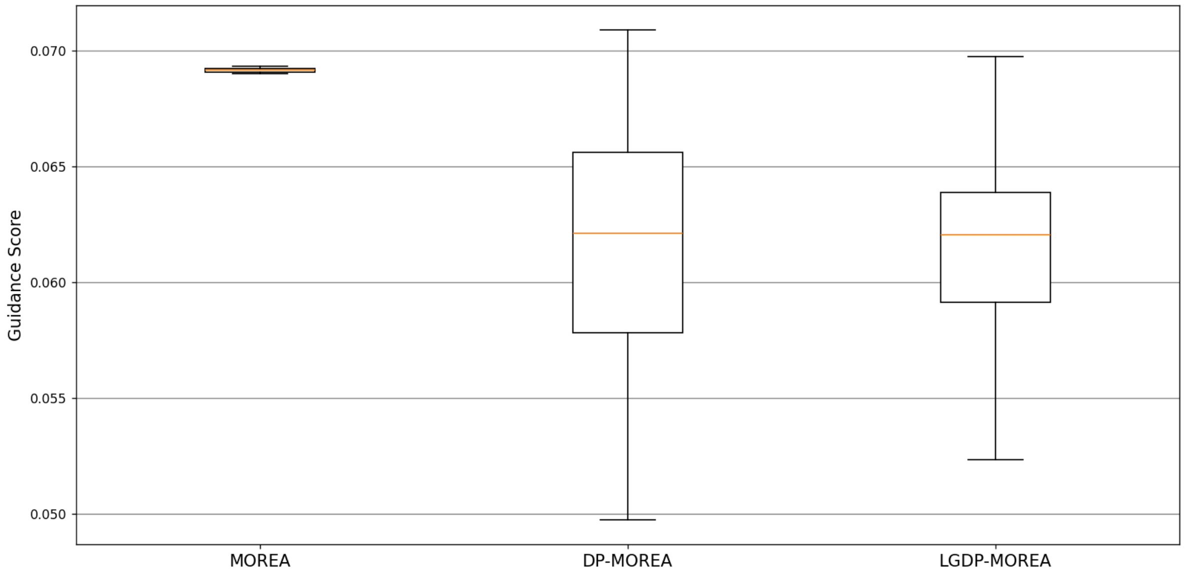
In Figures 5.13, 5.14, and 5.15, we have plotted inverse DVFs calculated from the solution in the initialized population with the lowest guidance value. There seem to be large differences in how the solutions are initialized. We observe that the DVFs of DP-MOREA are more homogeneous than MOREA and LGDP-MOREA. For all patients, the initializations of DP-MOREA and LGDP-MOREA, using DVFs instead of random deviations, seem to provide a good initial direction on where to start optimization.



(a) Patient 1

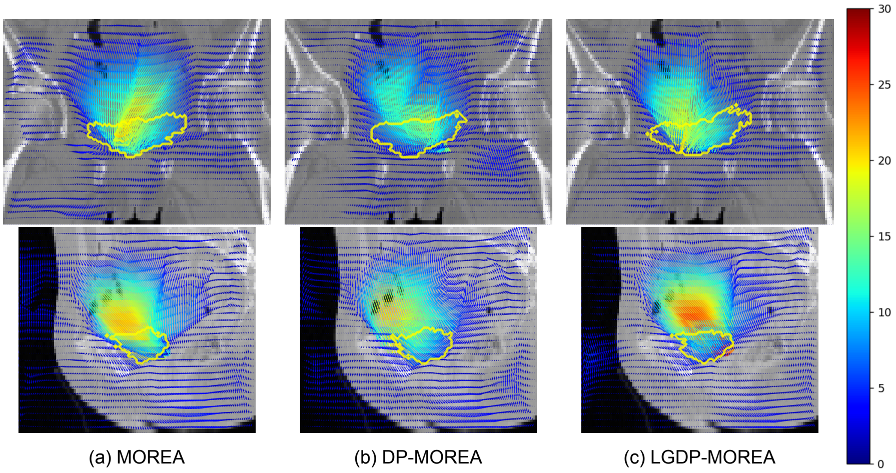


(b) Patient 2



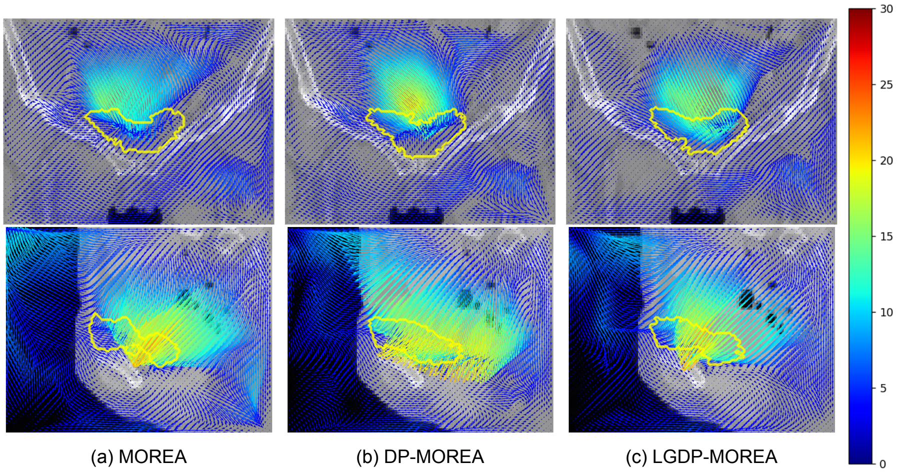
(c) Patient 3

Figure 5.12: Boxplots of the original MOREA objective scores for the initial population of MOREA, DP-MOREA and LGDP-MOREA for one repeat.



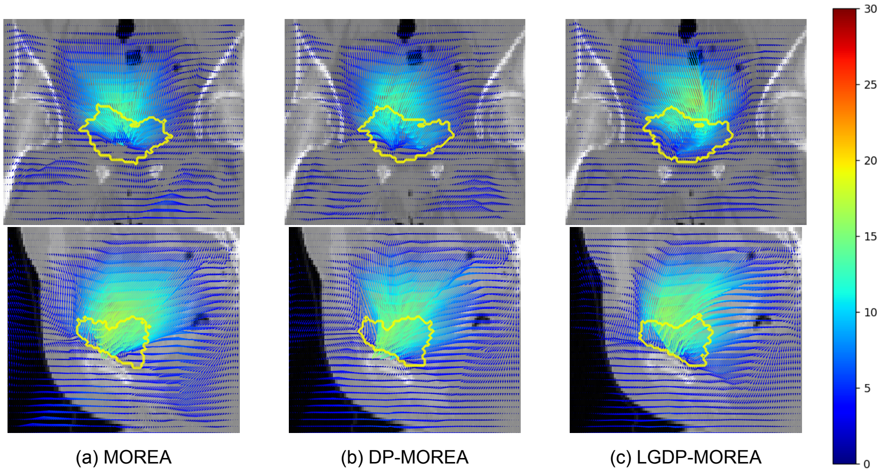
(a) MOREA (b) DP-MOREA (c) LGDP-MOREA

Figure 5.13: Downsampled DVFs and the deformed bladder contours of an initialized MOREA solution on Patient 1 in coronal (top) and sagittal (bottom) view. The arrow colors represent the vector magnitudes in voxel units.



(a) MOREA (b) DP-MOREA (c) LGDP-MOREA

Figure 5.14: Downsampled DVFs and the deformed bladder contours of an initialized MOREA solution on Patient 2 in coronal (top) and sagittal (bottom) view. The arrow colors represent the vector magnitudes in voxel units.



(a) MOREA (b) DP-MOREA (c) LGDP-MOREA

Figure 5.15: Downsampled DVFs and the deformed bladder contours of an initialized MOREA solution on Patient 3 in coronal (top) and sagittal (bottom) view. The arrow colors represent the vector magnitudes in voxel units.

5.3.2. Convergence of Minimal Guidance Objective

In Figure 5.16, convergence in terms of the minimal guidance objective of each generation of MOREA, DP-MOREA, and LGDP-MOREA is displayed for each repeat. In case the proposed DVF initialization method works, DP-MOREA and LGDP-MOREA should generally start with a lower minimum generation guidance objective than MOREA. Additionally, we expect convergence of this objective to be faster, as DP-MOREA and LGDP-MOREA both provide an informed initialization.

We observe that for both DP-MOREA and LGDP-MOREA, the minimum guidance score indeed starts at a lower value than MOREA. Next to this, each repeat of MOREA takes around 50 generations before the difference between minimum guidance scores in between generations increases, i.e., before convergence speeds up. DP-MOREA and LGDP-MOREA do not require this warm-up phase. As the generation number increases, MOREA converges on approximately the same minimal guidance scores as DP-MOREA and LGDP-MOREA, in the case of Patient 1 slightly surpassing both hybrids after approximately 175 generations. These observations indicate that the DVF initialization from the Digital Phantom approximation set does indeed provide an informed start for MOREA and has a positive effect on the convergence of the approach to high-quality solutions.

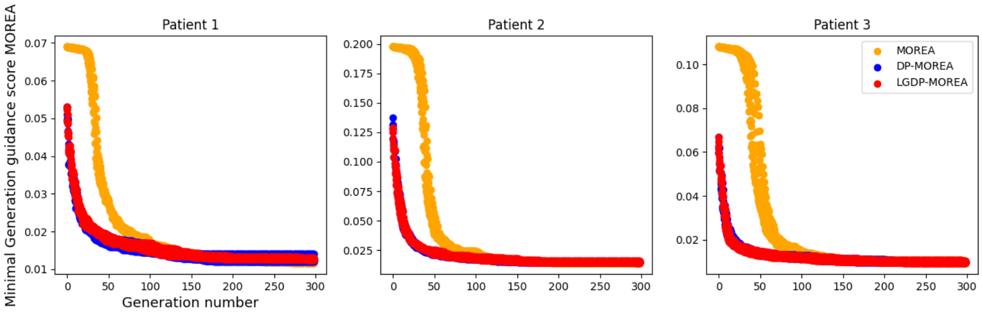


Figure 5.16: Convergence in terms of the minimal guidance objective per generation of all MOREA variants for all patients. All repeats are shown.

5.3.3. Hooke's Law and Hausdorff Distance Scores

In Figures 5.17, 5.18, and 5.19, the values for the Hooke's Law and Hausdorff Distance metric have been plotted for each DVF from the approximation set of MOREA, DP-MOREA, and LGDP-MOREA. Each line represents one repeat. Again, we expect to see a tradeoff between the Hooke's Law score and the Hausdorff Distance.

We observe that DP-MOREA and LGDP-MOREA are able to reach a lower Hausdorff Distance than MOREA for Patient 1. For Patient 2, all three approaches seem to have comparable performances for both metrics. For Patient 3, DP-MOREA and LGDP-MOREA reach lower Hooke's Law values for the same Hausdorff Distance values compared with MOREA. There is no clear indication of a better performance in either DP-MOREA or LGDP-MOREA. We can compare these findings with the hypervolume scores for these values, calculated with reference point ($[25, 10^7]$). These are depicted in Figures 5.20, 5.21, and 5.22. We observe that DP-MOREA and LGDP-MOREA seem to reach a higher hypervolume score than MOREA for two patients. The variance in hypervolume in between repeats is also similar for all approaches and patients.

Given the hypervolume, we are using a one-sided Mann-Whitney U test of DP-MOREA against MOREA and LGDP-MOREA against MOREA. Using this test result, we can determine whether the differences between these hybrids are significant. The results are displayed in Table 5.1. As mentioned in Chapter 4, we apply Bonferroni correction and reduce the α-level with a factor of 2 from 0.05 to 0.025.

The executed statistical tests indicate that both DP-MOREA and LGDP-MOREA are able to outperform MOREA in terms of the hypervolume in two out of three cases. Quantitatively, this indicates that the initialization information conveyed by the DVF approximation front can provide MOREA with higher Hooke's Law and Hausdorff Distance scores after the optimization loop.

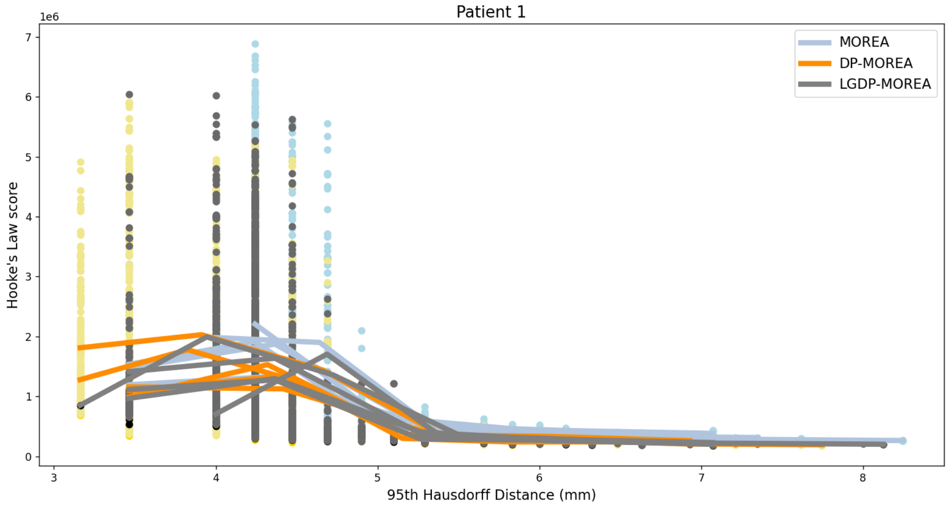


Figure 5.17: Calculated Hooke and Hausdorff values for the DVF approximation set of MOREA for Patient 1.

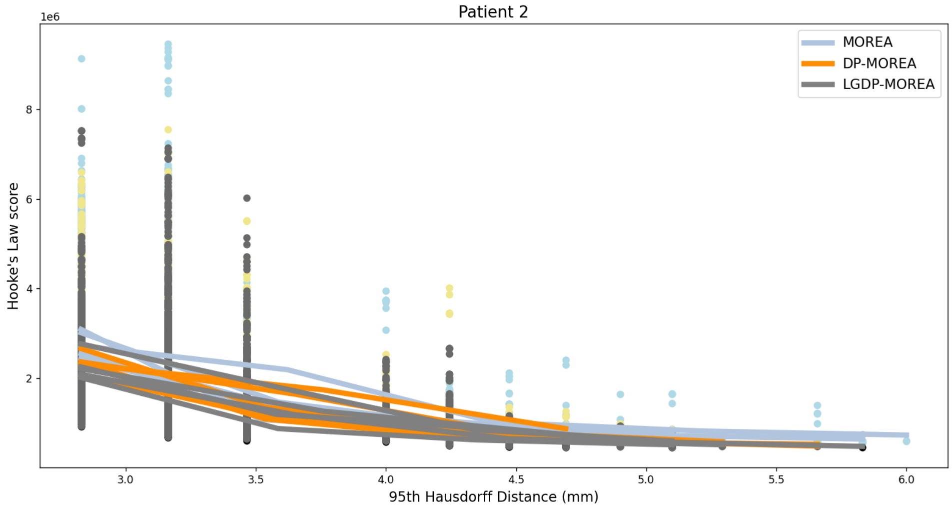


Figure 5.18: Calculated Hooke and Hausdorff values for the DVF approximation set of MOREA for Patient 2.

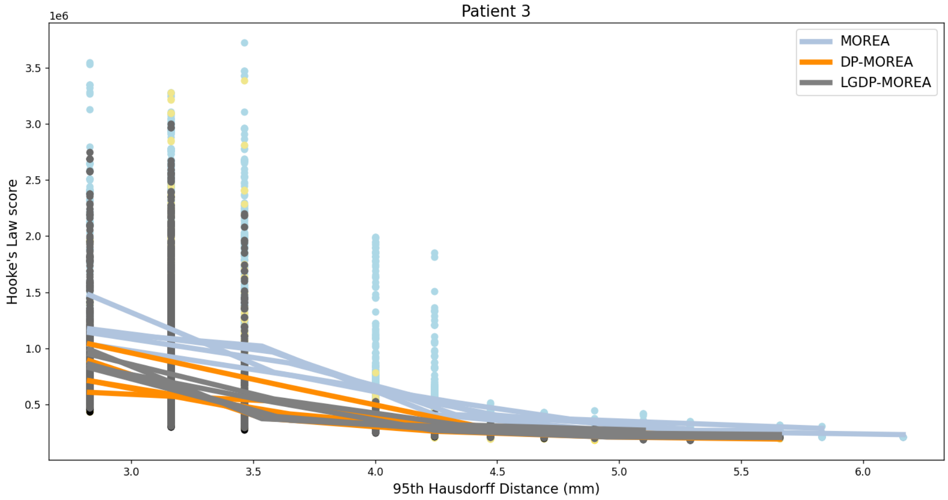


Figure 5.19: Calculated Hooke and Hausdorff values for the DVF approximation set of MOREA for Patient 3.

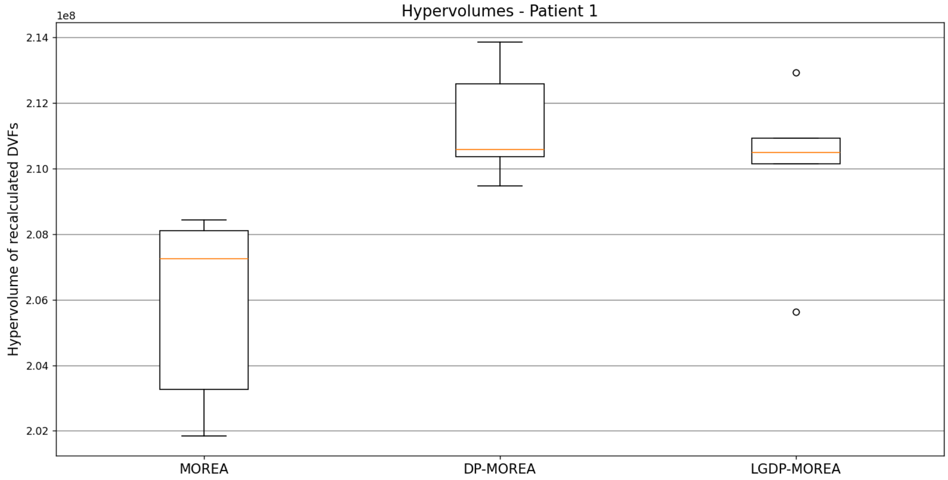


Figure 5.20: Box plots of MOREA Hypervolumes calculated from the Hooke's Law and Hausdorff scores. Each individual hypervolume represents one repeat on Patient 1.

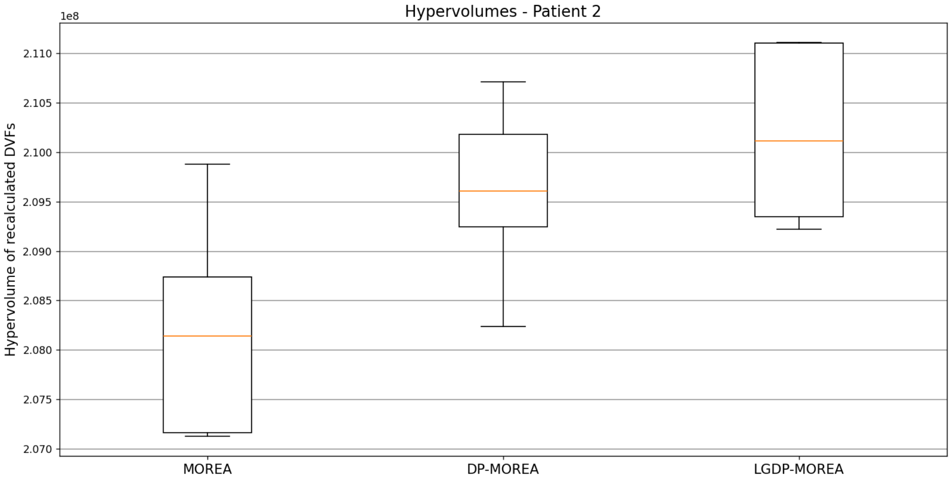


Figure 5.21: Box plots of MOREA Hypervolumes calculated from the Hooke's Law and Hausdorff scores. Each individual hypervolume represents one repeat on Patient 2.

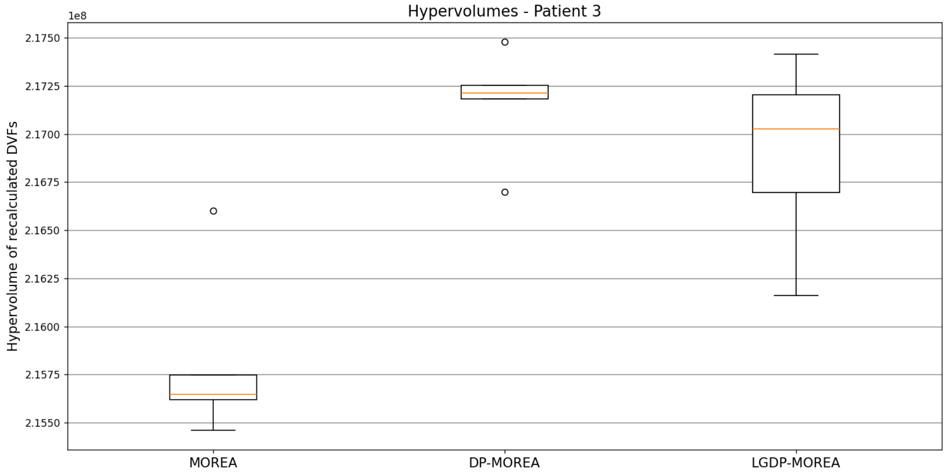


Figure 5.22: Box plots of MOREA Hypervolumes calculated from the Hooke's Law and Hausdorff scores. Each individual hypervolume represents one repeat on Patient 3.

Problem	DP-MOREA vs. MOREA	LGDP-MOREA vs. MOREA
Patient 1	**0.004**	0.028
Patient 2	0.048	**0.016**
Patient 3	**0.004**	**0.008**

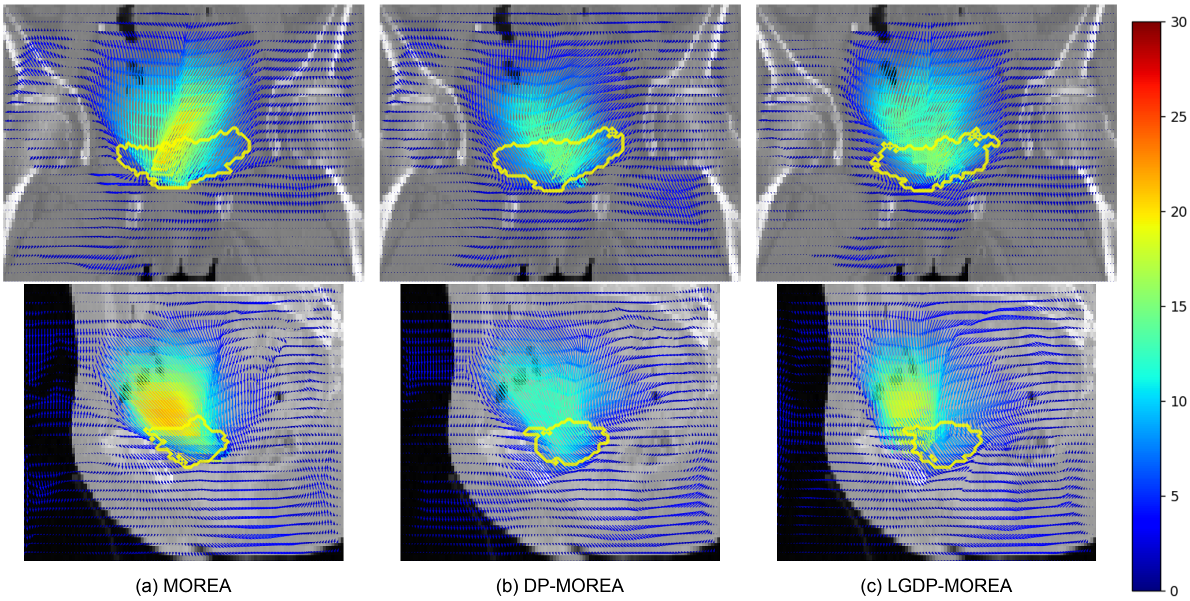
Table 5.1: p-values of the one-sided comparisons of hypervolumes based on Hooke's Law score and Hausdorff Distances of MOREA, DP-MOREA and LGDP-MOREA. Statistically significant results are highlighted in bold. DP-MOREA and LGDP-MOREA have better (higher) means in each case.

5.3.4. Qualitative Evaluation

Figures 5.23, 5.24, and 5.25, contain the same selected DVF slices for MOREA, DP-MOREA and LGDP-MOREA. In these DVFs, the largest magnitudes are centered around the bladder. In Figure 5.23, we observe that the DVF of DP-MOREA is more homogeneous compared to the other hybrids. MOREA and LGDP-MOREA seem to exhibit large deformation vector magnitudes directly around and in the bladder contours. In Figure 5.24, we notice the same trend. An interesting observation can be made for LGDP-MOREA, however. Here, the DVF magnitudes in the sagittal view abruptly change from a lower to a higher value inside the bladder contours. Apart from this, the DVFs seem about equal in homogeneity around the bladder contour.

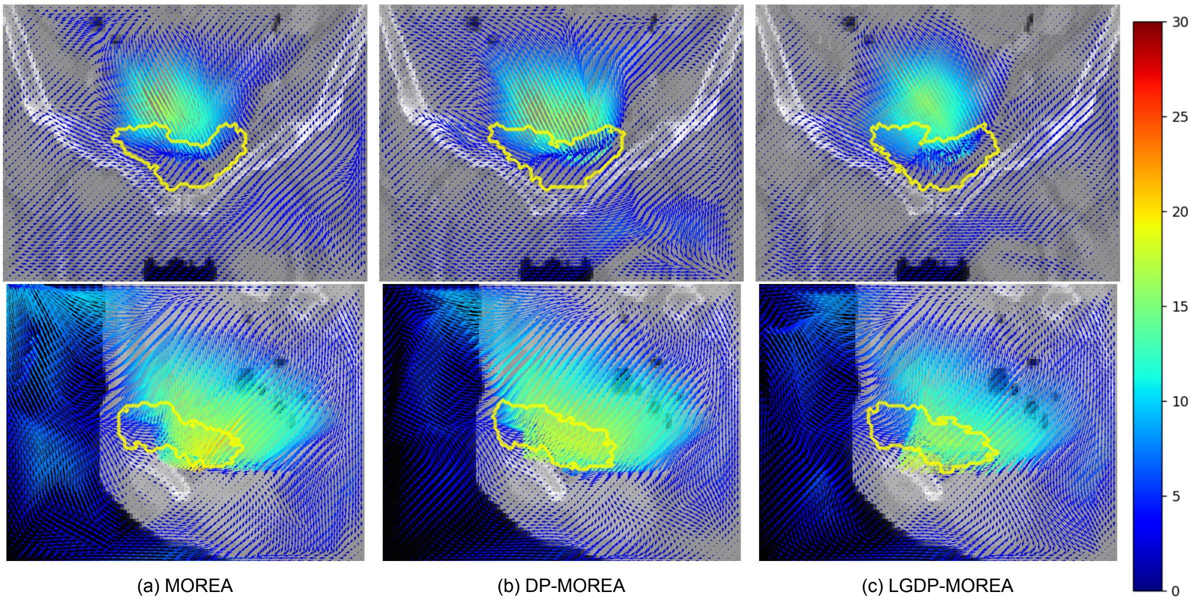
In Figure 5.25, we also notice the deformation magnitudes are mostly lower for DP-MOREA and LGDP-MOREA, which is especially visible in the sagittal view. The DVFs of DP-MOREA and LGDP-MOREA seem to be equally homogeneous, however. Again, the more homogeneous a DVF is, the fewer sudden displacements between neighbouring voxel vectors, which we consider a desirable characteristic. This could be explained by the way the DVF initialization method has been implemented. Since we effectively truncate the magnitudes of vectors when they result in a mesh fold, we are also homogenizing the deformation. Since this now more homogeneous mesh forms the initial population of MOREA, the approach is more likely to end up with more homogeneous solutions after optimization as well.

In Figure 5.26, the deformed bladder contours with the selected inverse DVFs are shown. We notice similar visual performances between all three hybrids. For Patient 1, LGDP-MOREA follows the desired bladder contours a little more closely. Apart from this, there are no observed large differences between the three hybrids. This indicates that qualitatively, there seem to be only small differences in contour fitting between MOREA, DP-MOREA, and LGDP-MOREA.



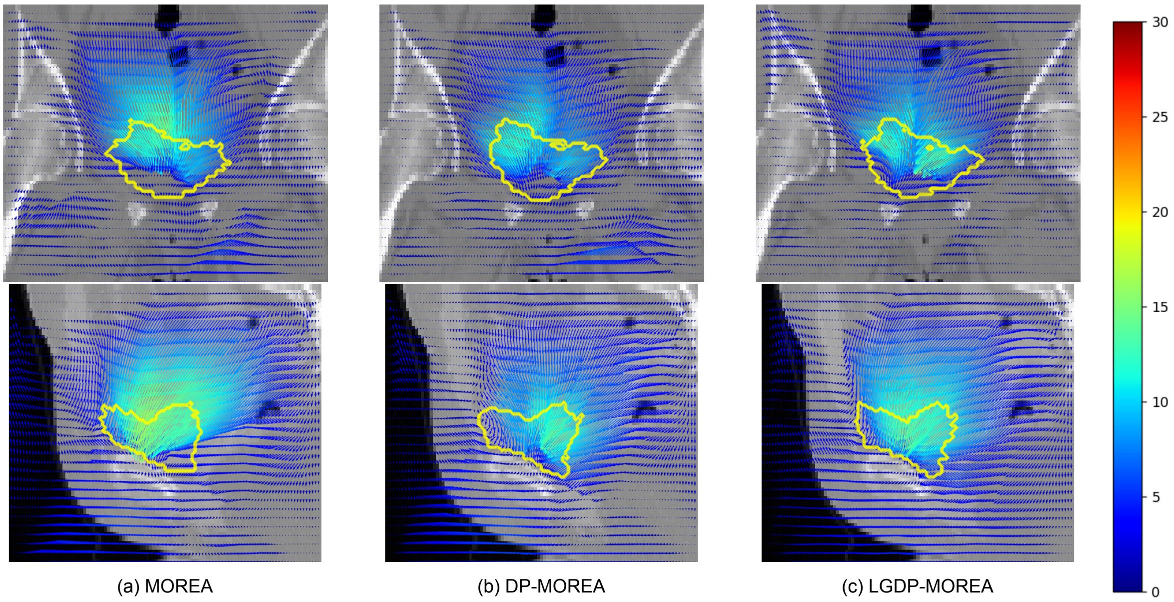
(a) MOREA (b) DP-MOREA (c) LGDP-MOREA

Figure 5.23: Patient 1: Downsampled DVFs and the deformed bladder contours of the MOREA configurations in coronal (top) and sagittal (bottom) view. The arrow colors represent the vector magnitudes in voxel units.



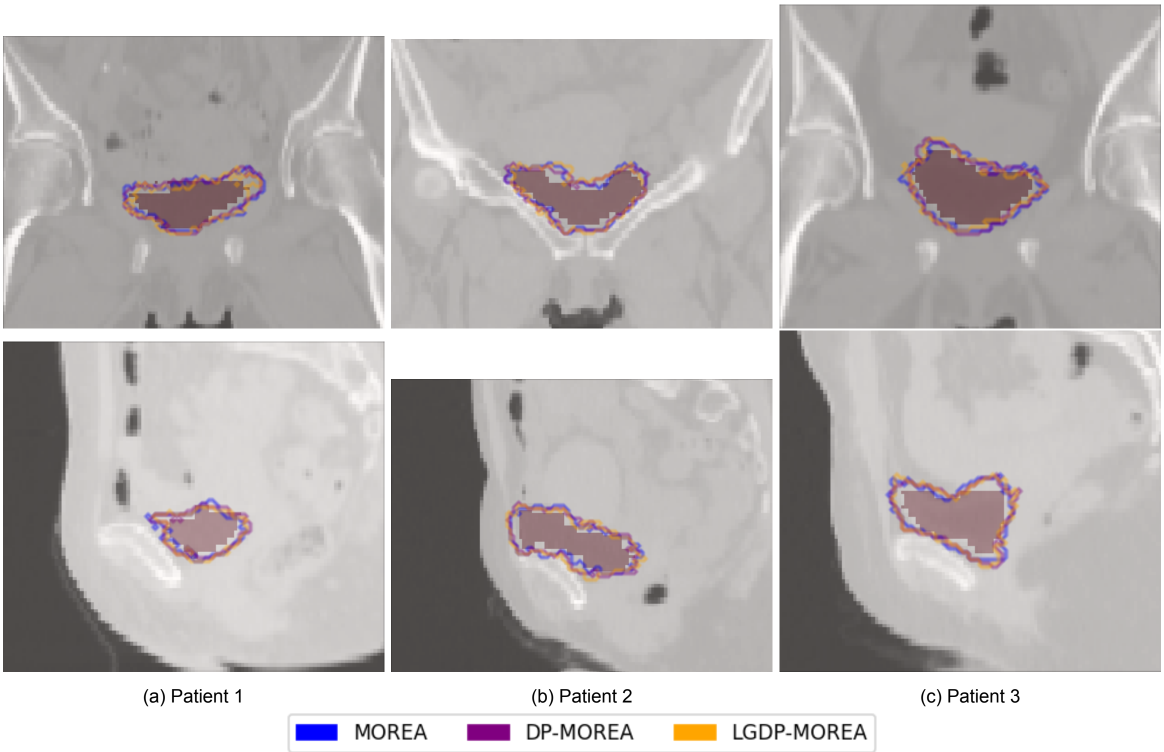
(a) MOREA (b) DP-MOREA (c) LGDP-MOREA

Figure 5.24: Patient 2: Downsampled DVFs and the deformed bladder contours of the MOREA configurations in coronal (top) and sagittal (bottom) view. The arrow colors represent the vector magnitudes in voxel units.



(a) MOREA (b) DP-MOREA (c) LGDP-MOREA

Figure 5.25: Patient 3: Downsampled DVFs and the deformed bladder contours of the MOREA configurations in coronal (top) and sagittal (bottom) view. The arrow colors represent the vector magnitudes in voxel units.



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

Figure 5.26: Deformed bladder contours on the source image for all patients and MOREA variants with the source bladder as a filled red area, in coronal (top) and sagittal (bottom) view.

5.4. Comparison of the Digital Phantom with MOREA, DP-MOREA, and LGDP-MOREA

In Figure 5.27, a box plot of the hypervolumes for all hybrids is depicted. We use the same reference point ($[25, 10^7]$) to calculate these hypervolumes. For Patients 1 and 2, we observe that the Digital Phantom and its hybrids are all outperformed in hypervolume by MOREA and its hybrids. For Patient 3, however, the baseline Digital Phantom appears to be similar to MOREA in performance. Simultaneously, its variance is much higher than the MOREA approaches, and the best obtained hypervolume still lies lower than the worst observed hypervolume of the MOREA hybrids. In Figure 5.28 we have depicted the deformed bladder contours of the MOREA variants and the baseline Digital Phantom. Due to the baseline Digital Phantom outperforming the LGDP configurations, we have opted to exclude those configurations from the full comparison.

For Patient 1 and 2, we notice that the deformed contours of MOREA, DP-MOREA, and LGDP-MOREA are a closer approximation of the desired bladder contours than the other approaches. For Patient 2, this is especially evident in the sagittal view. For Patient 3, we notice the desired bladder contours are followed much more closely by the deformed contours of the Digital Phantom than MOREA, DP-MOREA and LGDP-MOREA. This indicates that the Digital Phantom is able to find similar contours compared with MOREA, DP-MOREA and LGDP-MOREA. It also indicates that using the corresponding DVF from the Digital Phantom as initialization in MOREA can also lead to a degradation in contour match. We expect this might be due to the step-wise DVF initialization applied to MOREA, as the Digital Phantom mesh does contain small inverted volumes, which are not directly visible in the deformed contours.

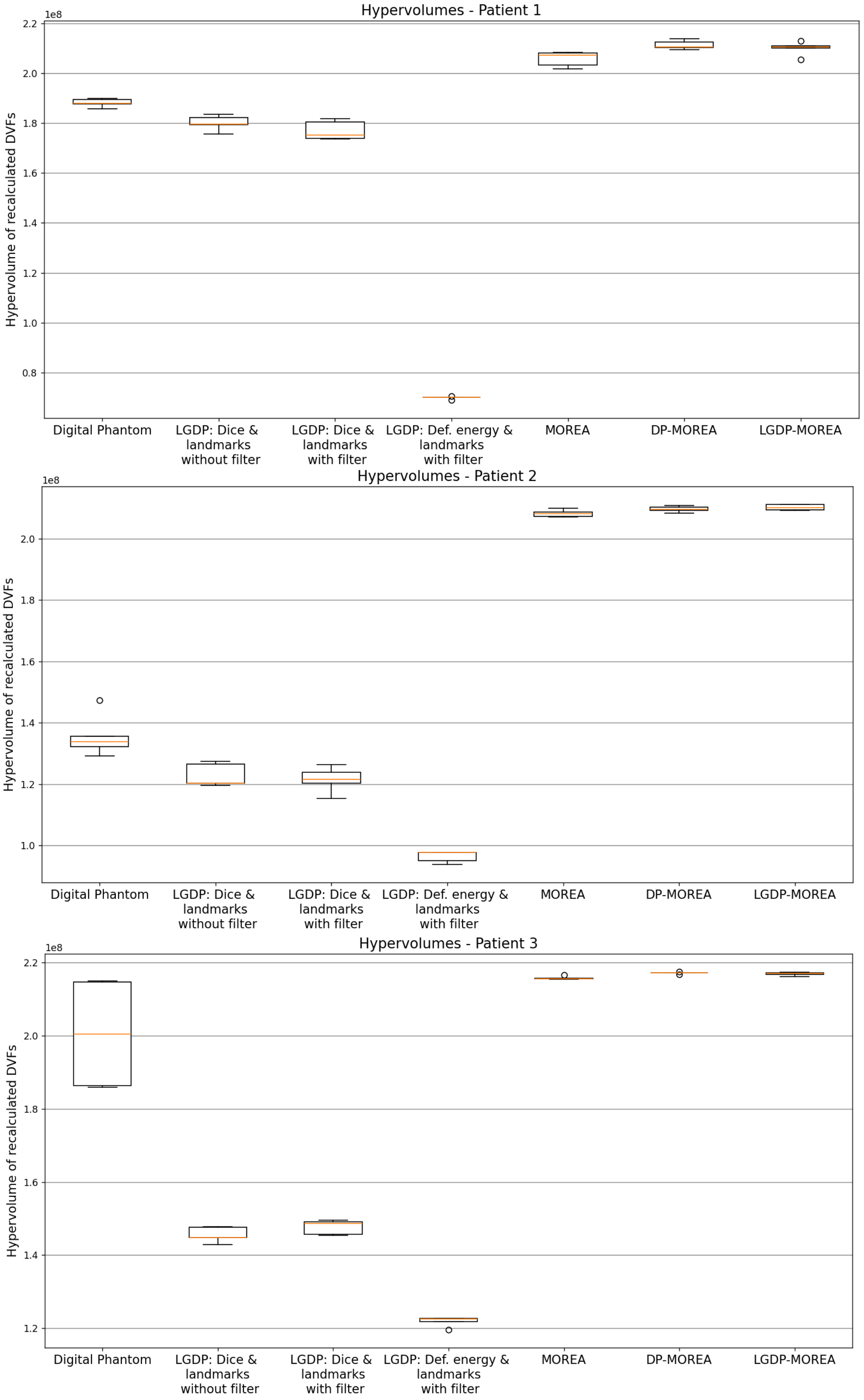


Figure 5.27: Boxplots of Digital Phantom and MOREA Hypervolumes calculated from the Hooke’s Law and Hausdorff scores. Each individual hypervolume represents one repeat.

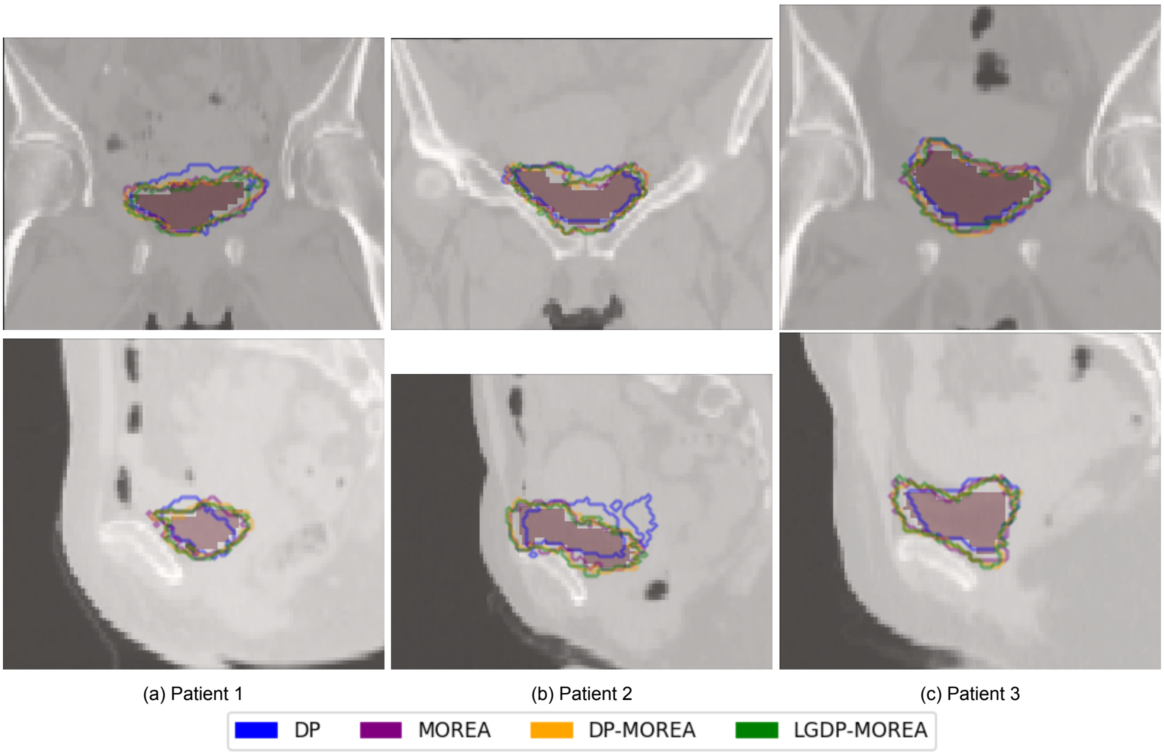

(a) Patient 1 (b) Patient 2 (c) Patient 3

Figure 5.28: Deformed bladder contours on the source image for all patients, MOREA variants and the baseline Digital Phantom with the reference bladder in full red in coronal (top) and sagittal (bottom) view.

6

Discussion

In this chapter, we discuss the results obtained in this work. Additionally, we discuss limitations in the investigation of the results and the created hybrids and mention possible future work directions.

6.1. Result Analysis

The results indicate that the landmarks obtained from the Automated Landmark Detection approach, in their current form, act more as a regularization method than a guidance objective. This is the case both for filtered and unfiltered landmarks. This regularization is especially evident from the LGDP Deformation Energy with Landmarks configuration, where barely any deformation occurs in the DVFs in the approximation front. This is also noticeable in the apparent tradeoff between the Dice and Landmark scores.

We see two potential reasons for these observations. Firstly, many landmark pairs describe very small to no displacements. Deforming tetrahedra containing these pairs will incur a distance penalty, making the guidance metric regularizing instead. Secondly, most landmarks are situated on the body and not directly in or around the bladder. This is due to the bladder volume being very small compared to the total volume of the body, allowing for less potential places for landmarks. Additionally, the region of the bladder has low intensity contrast on the source and target images, complicating corresponding landmark detection. Since we only optimized toward bladder deformation using the Dice score, these landmarks will not guide toward bladder deformation. Using landmark distance scores as a regularization objective, in their current form, has shown to result in large inverted volumes and infeasible meshes, in contrast to the baseline Digital Phantom, which finds less severe inversions. We conclude that the landmarks that describe much displacement do provide guidance information, but that the landmarks that remain stationary have a regularizing role.

Using landmarks does result in faster convergence of the original objectives of LGDP when the landmark correspondence objective is combined with the Dice score. The approximation front the LGDP configuration converges to is of lower quality than the approximation front of the Digital Phantom, however. Additionally, the Digital Phantom also results in qualitatively better fitting deformed contours than the LGDP configurations did. From this, we can conclude that when solely comparing the Digital Phantom with all the LGDP configurations, the baseline Digital Phantom outperforms the tested LGDP configurations both in the quantitative evaluation metrics and in deformation quality.

From the experiments regarding MOREA, DP-MOREA, and LGDP-MOREA, we can conclude that initializing the initial population of MOREA results in a faster improvement of the guidance objective of MOREA. Additionally, it results in higher guidance objective scores in the initial population. The DVF initialization leads to more variance in the parameter space and the objective space, in the initial population, compared to the standard initialization method. This results in a better exploration in the first stages of the optimization loop. Next to this, we conclude that both DP-MOREA and LGDP-MOREA show lower Hooke's Law and Hausdorff Distance values and more homogeneous DVFs than

MOREA. These experiments are backed up by a statistical test indicating the same finding. Qualitatively, MOREA, DP-MOREA, and LGDP-MOREA result in similarly fitting deformed contours. There is no clear evidence indicating that any single one of these hybrids consistently outperforms the others. We conclude that, although DP-MOREA and LGDP-MOREA have informed initialization, MOREA is able to qualitatively equal those hybrids in the end result.

In our comparison of the Digital Phantom with MOREA, DP-MOREA, and LGDP-MOREA, we observed that all three had better Hausdorff Distance values and hypervolumes for each patient than the Digital Phantom. Additionally, all three approaches had better fitting bladder contours for most patients. From this, we can conclude that further optimizing the approximation set from the Digital Phantom by using it as initialization of MOREA has a positive effect on the quality of its solutions within these evaluation metrics.

6.2. Limitations

There were some limitations in our investigation. Firstly, due to the limited scope of this master thesis, we reduced the problem from optimizing the deformation of all organs in the lower abdomen to only optimize the deformation of the bladder. Optimizing with more organs increases the complexity of the optimization problem, but could increase the usefulness of the landmark correspondences as guidance information.

The source meshes that we generated for the Digital Phantom approach were also reduced to a coarser mesh resolution. With a coarser mesh, the Digital Phantom's capability for capturing small, local deformations decreases significantly as there are fewer tetrahedra to deform. On the other hand, it does decrease the complexity of the problem, since a coarser mesh is more easily deformed. Modifying the mesh size influences the granularity of the inverse DVFs used in the evaluation. It could be possible that with a more detailed source mesh, the inverse DVFs are more fine-grained, which could result in higher evaluation scores and a more fair capability comparison.

A limitation to the quantitative evaluation in this work is that MOREA inherently already uses a distance metric as an objective. Since we use the Hausdorff Distance in our fair DVF comparison, this could give MOREA an advantage compared to the Digital Phantom, as the tests are inherently biased for contour distances. It could be possible that, with different evaluation metrics, the Digital Phantom has a higher general performance than MOREA. Additionally, our evaluation is based entirely on inverse DVFs. This is not a problem for MOREA, as the dual-dynamic mesh allows for accurate forward and inverse DVF estimation. The Digital Phantom, however, only optimizes the forward deformation using the FEM simulation. Calculating an inverse DVF from these solutions therefore introduces inaccuracies and does not represent the solution as well as a forward DVF would. Finally, these metrics do not represent the realism of the obtained deformation. It can be possible that the results are more realistic for the Digital Phantom than MOREA as it uses a FEM simulation. We tried to mitigate this possible misrepresentation by also qualitatively comparing the registrations.

We conducted a statistical test to compare MOREA with DP-MOREA and LGDP-MOREA. This test also has its limitations, as we only ran five repeats per hybrid and configuration. The choice of five repeats was made due to the time and resource limitations imposed on this thesis. Testing for statistical significance becomes more reliable as the amount of samples increases. Therefore, increasing the number of repeats could have made our statistical test more reliable. Additionally, this can provide further information on the variance between individual repeats.

Another limitation is the amount of clinical data available. We only used three patients in our evaluation. Using more patients, both with small and large differences in bladder volumes, can provide more insights into the actual performances of the approaches and hybrids.

Although these results look promising, the hybrids are still a long way from being used in clinical practice. Running the Digital Phantom, being a computationally expensive EA itself, to initialize MOREA adds a lot of computational expense. The time gain in terms of convergence speed induced by DP-

MOREA and LGDP-MOREA is therefore reduced by this additional processing. Next to this, MOREA is quantitatively able to catch up to DP-MOREA and LGDP-MOREA in terms of convergence, given more optimization time. Although we did not see large differences between the deformed contours of MOREA, DP-MOREA and LGDP-MOREA, using the results from the Digital Phantom could help MOREA find more realistic deformations. This is due to the possibility of the Digital Phantom being more realistic than MOREA as it uses a FEM simulation. This realism might not be directly visible when only considering deformed organ contours or the metrics used in our evaluation.

6.3. Future Work

The effect of the landmark filter on the quality of the DVFs and deformed contours of LGDP was not directly visible for the evaluated patients. More investigation should be performed on this filter. We recommend changing it to also filter out landmarks that exceed a distance to the nearest organ contour. Only considering landmarks that are close to organ contours could increase the guidance information with respect to these contours and reduce the regularizing aspect.

It may also prove useful to vary the source mesh granularity of the Digital Phantom. As mentioned, this could increase its local exploration capability and possibly result in more fine-grained deformations. Since higher-quality DVFs could also provide higher-quality initialization information for MOREA, this could provide more insight into its capabilities as an initialization method.

As we simplified our problem by reducing the number of optimizable organs, our comparison results are only applicable to this optimization setting. For future work, the impact of adding additional organs in the optimization process of the hybrids should be investigated. We expect that this could also further expose a difference between using initialization with a DVF approximation front and random initialization, as the gap in degrees of freedom between global and local exploration would increase.

We also encountered inverted volumes in the deformed meshes from the Digital Phantom. While in general it is desired to have meshes with no inverted volumes, it does not mean that a solution is directly invalid when it contains a single inversion. Still, we are looking for meshes that contain a minimal amount of these volumes, as they could introduce inaccuracies in the solution. We performed exploratory experiments by adding a hard inverted volume constraint during the optimization loop of the Digital Phantom, but this did not yield desirable results. Therefore, we recommend a more thorough exploration of the possibilities to prevent inverted volumes from occurring in the deformed meshes. This can, for example, be done by counting the inverted volumes for a solution during optimization and assigning the largest fitness value when this count exceeds a certain value.

The landmark correspondence metric was currently implemented by replacing either the Deformation Magnitude or the Dice objective of the Digital Phantom. It is also possible to introduce the landmark correspondence as a third objective instead. This could leverage both the guidance and regularizing capabilities of the landmarks, but also optimize with respect to the contour similarities and the global regularization of the deformation magnitude. Additionally, the landmarks could be considered directly as an objective in MOREA. This can also come with its own various hybrids.

The landmarks that we used were automatically detected using deep learning and might not all be accurate. Investigating the impact of using landmarks in the Digital Phantom and MOREA can be further extended by using manually annotated landmarks instead. We expect these landmarks can be more accurate and thus provide better guidance information for the approaches.

7

Conclusion

In this work, we have investigated the impact of creating hybrid versions of two methods for Deformable Image Registration: the Digital Phantom (Rodriguez et al., 2023) and MOREA (Andreadis et al., 2023). In this process, we also analyzed and used landmarks from an Automated Landmark Detection approach (Grewal et al., 2023). To ensure these landmarks have high quality, we designed a filtering strategy. These landmarks were used in the Digital Phantom as a guidance objective. This was done with various configurations, resulting in the Landmark Guided Digital Phantom hybrid. Furthermore, we implemented a way to initialize MOREA with the results of the Digital Phantom, resulting in Digital Phantom MOREA (DP-MOREA) and the Landmark Guided Digital Phantom MOREA (LGDP-MOREA). We tested these hybrid methods on clinical scans, featuring deformations of the bladder of three cervical cancer patients. To ensure a fair comparison in our quantitative evaluation, we evaluated the performances using a Hooke's Law-based score, the Hausdorff Distance and a hypervolume score based on these metrics. We also qualitatively evaluated the solutions by performing a visual analysis of the resulting Deformation Vector Fields and deformed bladder contours.

The results showed that the landmarks, which were initially believed to provide a guiding role in optimization, instead worked in a regularizing capacity. Using the landmarks as a regularizing objective in the Digital Phantom instead of the original Deformation Energy objective yielded worse results. We concluded that the Deformation Energy is best suited for global regularization. The landmarks both regularize and provide guidance information, thus forming a less effective regularization method. The Digital Phantom was thus found to be outperforming LGDP on the tested patients.

Using the results from the Digital Phantom to initialize MOREA yielded satisfactory results. The initial population of MOREA contained more diverse solutions and was already initialized in the direction of the optimal deformation. This resulted in faster convergence of DP-MOREA and LGDP-MOREA compared to MOREA. We concluded that DP-MOREA and LGDP-MOREA both outperform MOREA, quantitatively. Qualitatively, the end results of MOREA do not differ much from those of DP-MOREA and LGDP-MOREA.

We finalize this work by mentioning that even though DP-MOREA and LGDP-MOREA are able to outperform the original algorithms and hybrids, they are still not ready to be used in clinical practice. Creating the informed initialization of MOREA requires a lot of computational resources, diminishing the effect of the initial faster convergence. The results do show promise, however, as initializing MOREA using the Digital Phantom could still result in more realistic deformations. We find that more research is required to investigate the differences between MOREA, DP-MOREA, and LGDP-MOREA and their clinical relevance.

A

Appendix A

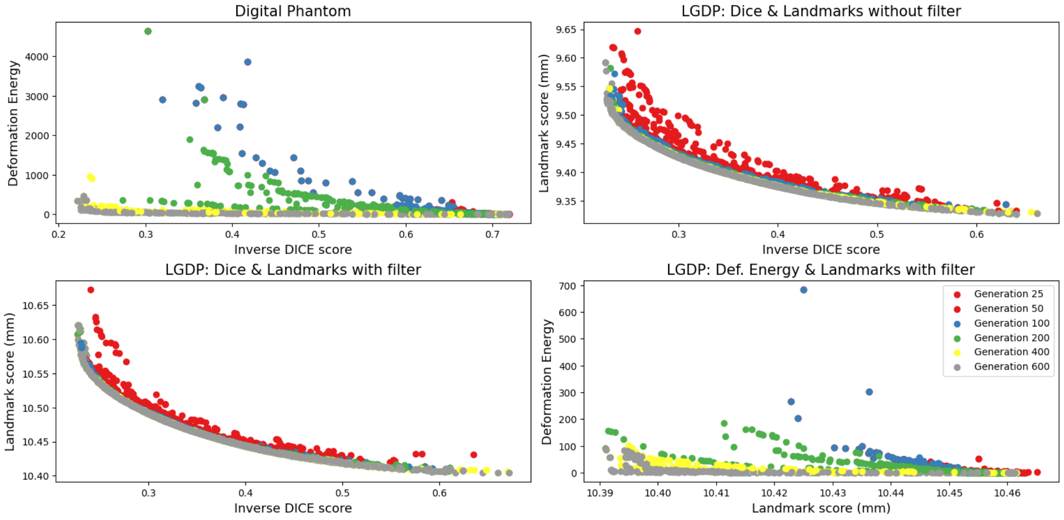


Figure A.1: Convergence in terms of the original objectives of the Digital Phantom for Patient 2.

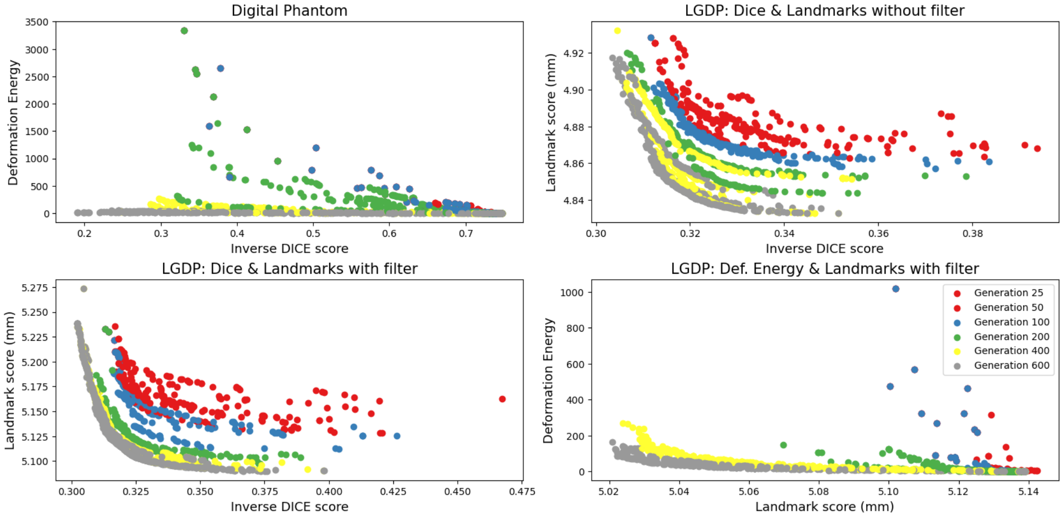


Figure A.2: Convergence in terms of the original objectives of the Digital Phantom for Patient 3.

Bibliography

Taylor, A., & Powell, M. E. (2008). An assessment of interfractional uterine and cervical motion: Implications for radiotherapy target volume definition in gynaecological cancer. *Radiotherapy and oncology*, *88*(2), 250–257.

Oh, S., & Kim, S. (2017). Deformable image registration in radiation therapy. *Radiation oncology journal*, *35*(2), 101.

Andreadis, G., Bosman, P. A., & Alderliesten, T. (2023). Morea: A gpu-accelerated evolutionary algorithm for multi-objective deformable registration of 3d medical images. *arXiv preprint arXiv:2303.04873*.

Sotiras, A., Davatzikos, C., & Paragios, N. (2013). Deformable medical image registration: A survey. *IEEE Transactions on Medical Imaging*, *32*(7), 1153–1190. https://doi.org/10.1109/TMI.2013.2265603

Pirpinia, K., Bosman, P. A., Loo, C. E., Winter-Warnars, G., Janssen, N. N., Scholten, A. N., Sonke, J.-J., Van Herk, M., & Alderliesten, T. (2017). The feasibility of manual parameter tuning for deformable breast mr image registration from a multi-objective optimization perspective. *Physics in Medicine & Biology*, *62*(14), 5723.

Alderliesten, T., Bosman, P. A., & Bel, A. (2015). Getting the most out of additional guidance information in deformable image registration by leveraging multi-objective optimization. *Medical Imaging 2015: Image Processing*, *9413*, 469–475.

Rigaud, B., Simon, A., Castelli, J., Lafond, C., Acosta, O., Haigron, P., Cazoulat, G., & de Crevoisier, R. (2019). Deformable image registration for radiation therapy: Principle, methods, applications and evaluation. *Acta Oncologica*, *58*(9), 1225–1237.

Rodriguez, C., de Boer, S., Bosman, P., & Alderliesten, T. (2023). Bi-objective optimization of organ properties for the simulation of intracavitary brachytherapy applicator placement in cervical cancer, 16. https://doi.org/10.1117/12.2647129

Grewal, M., Deist, T. M., Wiersma, J., Bosman, P. A. N., & Alderliesten, T. (2020). An end-to-end deep learning approach for landmark detection and matching in medical images [Landmark detection and comparison using Deep Learning]. https://doi.org/10.1117/12.2549302

Weistrand, O., & Svensson, S. (2015). The anaconda algorithm for deformable image registration in radiotherapy. *Medical physics*, *42*(1), 40–53.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, *26*(3), 297–302.

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

Deb, K. (2011). Multi-objective optimisation using evolutionary algorithms: An introduction. In L. Wang, A. H. C. Ng, & K. Deb (Eds.), *Multi-objective evolutionary optimisation for product design and manufacturing* (pp. 3–34). Springer London. https://doi.org/10.1007/978-0-85729-652-8_1

Cheng, R., Jin, Y., Olhofer, M., & Sendhoff, B. (2016). A reference vector guided evolutionary algorithm for many-objective optimization. *IEEE Transactions on Evolutionary Computation*, *20*(5), 773–791. https://doi.org/10.1109/TEVC.2016.2519378

Deb, K., Agrawal, R. B., et al. (1995). Simulated binary crossover for continuous search space. *Complex systems*, *9*(2), 115–148.

Bouter, A., Luong, N. H., Witteveen, C., Alderliesten, T., & Bosman, P. A. (2017). The multi-objective real-valued gene-pool optimal mixing evolutionary algorithm. *Proceedings of the Genetic and Evolutionary Computation Conference*, 537–544.

Floater, M. S. (2015). Generalized barycentric coordinates and applications. *Acta Numerica*, *24*, 161–214.

Bosman, P. A. (2010). The anticipated mean shift and cluster registration in mixture-based edas for multi-objective optimization. *Proceedings of the 12th Annual Conference on Genetic and Evolutionary Computation*, 351–358. https://doi.org/10.1145/1830483.1830549

Huebner, K. H., Dewhirst, D. L., Smith, D. E., & Byrom, T. G. (2001). *The finite element method for engineers*. John Wiley & Sons.

Haskins, G., Kruger, U., & Yan, P. (2020). Deep learning in medical image registration: A survey. *Machine Vision and Applications*, *31*, 1–18.

Kearney, V., Chen, S., Gu, X., Chiu, T., Liu, H., Jiang, L., Wang, J., Yordy, J., Nedzi, L., & Mao, W. (2014). Automated landmark-guided deformable image registration. *Physics in Medicine & Biology*, *60*(1), 101.

Yang, D., Zhang, M., Chang, X., Fu, Y., Liu, S., Li, H. H., Mutic, S., & Duan, Y. (2017). A method to detect landmark pairs accurately between intra-patient volumetric medical images. *Medical physics*, *44*(11), 5859–5872.

Grewal, M., Wiersma, J., Westerveld, H., Bosman, P. A. N., & Alderliesten, T. (2023). Automatic landmark correspondence detection in medical images with an application to deformable image registration. *Journal of Medical Imaging*, *10*. https://doi.org/10.1117/1.jmi.10.1.014007

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. http://arxiv.org/abs/1505.04597

Andreadis, G., Mulder, J. I., Bouter, A., Bosman, P. A., & Alderliesten, T. (2024). A tournament of transformation models: B-spline-based vs. mesh-based multi-objective deformable image registration. *arXiv preprint arXiv:2401.16867*.

Huttenlocher, D. P., Klanderman, G. A., & Rucklidge, W. J. (1993). Comparing images using the hausdorff distance. *IEEE Transactions on pattern analysis and machine intelligence*, *15*(9), 850–863.

Guerreiro, A. P., Fonseca, C. M., & Paquete, L. (2021). The hypervolume indicator: Computational problems and algorithms. *ACM Computing Surveys (CSUR)*, *54*(6), 1–42.