

Variational Methods in Shape Space

Dissertation
zur Erlangung des Doktorgrades (Dr. rer. nat.)
der Mathematisch-Naturwissenschaftlichen Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von
Benedikt Konstantin Josef Wirth
aus Kiel

Bonn, November 2009

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der
Rheinischen Friedrich-Wilhelms-Universität Bonn
am Institut für Numerische Simulation

1. Gutachter: Prof. Dr. Martin Rumpf
2. Gutachter: Prof. Dr. Stefan Müller
3. Gutachter: Prof. Dr. David Mumford

Tag der Promotion: 2. Juni 2010
Erscheinungsjahr: 2010

Zusammenfassung

Die vorliegende Arbeit befasst sich mit Anwendungen von Variationsmethoden in Räumen geometrischer Formen. Insbesondere werden die Mittelung und Hauptkomponentenanalyse von Formen, die Berechnung geodätischer Pfade zwischen zwei Formen sowie Formoptimierung behandelt.

Einen kurzen Überblick über die zugrunde gelegte Modellierung des Formenraums gibt Kapitel 1. Geometrische Formen werden mit zwei- oder dreidimensionalen, deformierbaren Körpern oder Objekten identifiziert. Zur Beschreibung der Deformationen dienen physikalische Modelle; im Speziellen wird entweder von einem hyperelastischen oder von einem viskosen Materialverhalten der Objekte ausgegangen. Die insbesondere für eine numerische Implementierung der Variationsansätze wichtige Darstellung der Objekte mit Hilfe von Phasenfeldern oder Niveaumengen von Funktionen wird ebenfalls kurz dargelegt.

Kapitel 2 enthält eine Übersicht über unterschiedliche sowie mit dieser Arbeit verwandte Ansätze der Modellierung von Formenräumen. Des Weiteren werden Referenzen zu verschiedenen Bildsegmentierungs- und -registrierungsmethoden angegeben, auf die an späterer Stelle zurückgegriffen wird. Schließlich wird die in dieser Arbeit relevante Literatur zur Formoptimierung vorgestellt.

Kapitel 3 umfasst eine Einführung in die benötigten Konzepte aus der Kontinuumsmechanik und der Phasenfeldmodellierung. Insbesondere werden einige in den folgenden Kapiteln angewandte Sätze angegeben.

Kapitel 4 beschäftigt sich mit der Durchschnittsbildung geometrischer Formen, basierend auf einem hyperelastischen Abstands begriff: Die Ähnlichkeit einer Form zu einer anderen wird gemessen als minimal erforderliche Deformationsenergie, um die erste Form in die zweite zu überführen. Es wird ein entsprechendes Phasenfeldmodell entworfen, analysiert und anschließend numerisch mittels Finiten Elemente implementiert.

Eine mit dem hyperelastischen Abstands begriff konsistente Hauptkomponentenanalyse gegebener Formen wird in Kapitel 5 behandelt. Elastische Spannungen auf der gemittelten Form dienen dabei als Repräsentanten der Eingangsformen in einem linearen Vektorraum. Auf diesen linearen Repräsentanten kann eine klassische Hauptkomponentenanalyse durchgeführt werden. Hierbei sollte das Skalarprodukt des Vektorraums passend zu den Eingangsformen gewählt werden und nicht unabhängig von ihnen sein.

In Kapitel 6 werden geometrische Formen als Objekte aus viskosem Material modelliert, um geodätische Pfade zwischen ihnen zu definieren. Die Länge eines Pfades ist dabei gegeben als die gesamte physikalische Dissipation, die während der Deformation eines Objektes entlang des Pfades entsteht. Eine zeitliche Diskretisierung dieser Pfade, die invariant bezüglich Starrkörperbewegungen ist, wird erreicht, indem die Dissipation entlang eines Pfadsegments durch die elastische Deformationsenergie eines elastisch deformierten Objekts approximiert wird. Auf dieser Grundlage wird schließlich eine numerische Implementierung mit Hilfe von Niveaumengen-Funktionen vorgenommen.

Kapitel 7 befasst sich mit der Formoptimierung elastisch beanspruchter Strukturen.

Für vorgegebene mechanische Lasten soll die Balance zwischen Starrheit der Struktur, ihrem Volumen und ihrer Oberfläche optimiert werden. Zu diesem Zweck wird ein Phasenfeldmodell aufgestellt und analysiert. Die Nutzung nichtlinearer Elastizität erlaubt hierbei, sogenannte Knickfälle zu erkennen, die bei Benutzung linearisierter Elastizität ignoriert werden.

Teile dieser Arbeit wurden bereits in Fachzeitschriften und auf Konferenzen veröffentlicht [104, 103, 127, 105, 106] beziehungsweise eingereicht [128, 96].

Contents

1	Overview	1
1.1	Shapes as boundaries of deformable objects	1
1.2	Shape representations	3
1.3	Coupling of deformations and shape representations	4
2	Related work on shape spaces and shape optimisation	7
2.1	Different approaches to shape space	7
2.2	Registration	12
2.3	Segmentation	12
2.4	Shape optimisation	13
3	Concepts from elasticity and multiphase modelling	21
3.1	Solid continuum mechanics	21
3.2	Phase field description of free interface and discontinuity problems	27
4	Elastic shape averaging	31
4.1	Shape averages based on nonlinear elastic distances	32
4.1.1	Variational definition of the shape average	33
4.1.2	A relaxed energy formulation	36
4.1.3	Joint averaging and segmentation	37
4.2	Phase field approximation	38
4.2.1	Statement of the averaging energy in terms of phase fields	39
4.2.2	Euler–Lagrange equations	41
4.2.3	Existence of minimisers	43
4.3	Numerical implementation	46
4.4	Validation and experiments	50
4.4.1	Averaging of 2D shapes	50
4.4.2	Averaging of 3D shapes	51
4.4.3	Weighted averaging	54
4.4.4	Averaging image morphologies	55
4.5	Discussion	57
5	Principal modes of elastic shape variation	59
5.1	Linearisation of shape variations	61
5.1.1	The associated mechanical problem	61

5.1.2	Boundary stresses as alternative linearisations	64
5.2	Covariance analysis	65
5.2.1	Impact of the covariance metric	68
5.2.2	Impact of the nonlinear elastic constitutive law	68
5.2.3	Elastic versus Riemannian shape analysis	69
5.3	Numerical implementation	72
5.4	Validation and applications	73
5.4.1	PCA for 2D input shapes	73
5.4.2	PCA for image morphologies	75
5.4.3	PCA for 3D shapes	75
6	Geodesics based on viscous flow and their variational time discretisation	79
6.1	Paths in shape space generated by viscous deformation	80
6.2	Variational time discretisation	82
6.2.1	Discrete geodesics	85
6.2.2	A relaxed formulation	88
6.2.3	Viscous fluid model for vanishing time step size	89
6.3	Regularised level set approximation	92
6.4	Numerical implementation	94
6.5	Examples and generalisations	96
6.5.1	A fragment of shape space	97
6.5.2	Influence of material parameters	100
6.5.3	Geodesics between partially occluded shapes	101
6.5.4	Geodesics between multilabelled images	101
6.5.5	Shape clustering via geodesic distances	105
6.6	Discussion	105
7	Minimum compliance design in nonlinear elasticity	107
7.1	Geometry optimisation for a prescribed mechanical load	108
7.1.1	Balancing compliance with volume and perimeter	109
7.1.2	Allen–Cahn phase field approximation	110
7.2	Effect of nonlinear elasticity in shape optimisation	111
7.2.1	Choice of compliance definition	111
7.2.2	Proper handling of load asymmetries	113
7.2.3	Buckling instabilities	114
7.3	Model analysis	118
7.3.1	Existence of minimisers	118
7.3.2	Non-existence of minimisers for worst case deformations	121
7.3.3	Model behaviour for phase field parameter $\varepsilon \rightarrow 0$	124
7.4	Numerical implementation	126
7.4.1	Optimality conditions and finite element discretisation	127

7.4.2	Inner minimisation to find equilibrium deformation	128
7.4.3	Optimisation for the phase field	130
7.4.4	Embedding the optimisation in a multiscale approach	131
7.5	Experiments	131
Bibliography		143

1 Overview

This thesis deals with some variational problems in the space of elastically or viscously deformable objects or shapes. As topics of central importance to the analysis of a shape space we will examine the problem of defining and computing an average for a given set of shapes in Chapter 4, the extraction of dominant modes of shape variation in Chapter 5, the computation of geodesics between shapes in Chapter 6, and shape optimisation under PDE-constraints in Chapter 7. The following sections will provide a brief overview over the employed notion of shapes as well as the techniques to represent them.

1.1 Shapes as boundaries of deformable objects

There are many possible approaches to define a space of shapes. As examples, consider the vector space of landmark positions, where each shape corresponds to a vector of points in \mathbb{R}^d , or a space of sufficiently regular subsets of \mathbb{R}^d , endowed with an appropriate metric (for instance, the Hausdorff distance or some measure of the symmetric difference between sets), or the space of level set functions $u : \mathbb{R}^d \rightarrow \mathbb{R}$ with the structure induced, for example, by the vector space of real-valued, continuous functions on \mathbb{R}^d .

In this work, we will desire some specific shape space properties. In particular, we would like to be able to properly describe large and strongly nonlinear geometric variations of shapes, and we require our shape space to be rigid body motion invariant, that is, rotated and translated versions of a shape shall be identified with each other.

A physically intuitive idea of shapes, which is able to fulfil the above properties, is to regard shapes \mathcal{S} as the boundaries $\partial\mathcal{O}$ of deformable objects \mathcal{O} . Following this idea, we will understand different elements of the shape space as deformed configurations of each other. The shape space structure then translates into the structure of the space of deformations between the shapes or rather between the corresponding objects. Note that this notion of shapes differs from a purely geometric interpretation as planar curves or 3D surfaces: We associate with each shape \mathcal{S} an object \mathcal{O} which represents its inside and which undergoes deformations if the shape changes.

There are various possibilities to impose a structure on such a shape space. Throughout this work, this structure will be based on physical deformation energy in the sense that we will mimic deformations of real physical objects when describing the relation between different shapes. The physical energy required to deform one object \mathcal{O} into another will serve as a measure of how close the corresponding shapes are. This already implies rigid body motion invariance and the tendency to locally preserve isometry, since

translations or rotations of objects cost zero energy. Since we would like to allow large deformations, there are two natural concepts that can be applied: *Nonlinear elasticity* and *viscous flow*.

In the case of nonlinear elasticity, the distance from a shape $\mathcal{S}_1 = \partial\mathcal{O}_1$ to $\mathcal{S}_2 = \partial\mathcal{O}_2$ will be based on the nonlinear elastic deformation energy of the energetically optimal deformation $\phi : \mathcal{O}_1 \rightarrow \mathbb{R}^d$ with $\phi(\mathcal{O}_1) = \mathcal{O}_2$. This energy can (under certain assumptions about the modelled physical material) be expressed in the form

$$\mathcal{W}[\mathcal{O}_1, \phi] = \int_{\mathcal{O}_1} \hat{W}(\|\mathcal{D}\phi\|_F, \|\text{cof}\mathcal{D}\phi\|_F, \det\mathcal{D}\phi) \, dx,$$

where $\mathcal{D}\phi$ represents the deformation gradient, $\text{cof}\mathcal{D}\phi = \det\mathcal{D}\phi \mathcal{D}\phi^{-T}$ denotes its cofactor matrix and $\|\mathcal{D}\phi\|_F = \sqrt{\text{tr}(\mathcal{D}\phi^T \mathcal{D}\phi)}$ its Frobenius norm (see Section 3.1).

In this elastic setting, there is no natural notion of paths between two shapes or objects: Due to the fundamental axiom of elasticity, the energy of a deformation is completely independent of the path along which this final deformation was reached. Also, the deformation energy from one shape to another (and also its square root) is in general neither symmetric (indeed, we normally have $\mathcal{W}[\mathcal{O}, \phi] \neq \mathcal{W}[\phi(\mathcal{O}), \phi^{-1}]$), nor does it satisfy the triangle inequality: Usually, deforming an object \mathcal{O}_1 into an object \mathcal{O}_2 costs more energy than the sum of deformation energies of deforming \mathcal{O}_1 into some intermediate object $\tilde{\mathcal{O}}$ and $\tilde{\mathcal{O}}$ into \mathcal{O}_2 . Consequently, nonlinear elastic deformation energy will not induce a metric distance. Hence, this concept is particularly appropriate when we are interested in unidirectional comparisons of a number of shapes with one single distinguished shape, for example, if we try to construct a single deformable template for a whole set of shapes. For this reason, we will use elastic distances for shape averaging in Chapter 4.

When using the concept of viscous deformations, the distance between two shapes or objects will be based on the viscous dissipation during the deformation of one into the other. This approach imposes a Riemannian structure on the shape space: The tangent space at a shape $\partial\mathcal{O}$ comprises all velocity fields on \mathcal{O} , and its first fundamental form (for material parameters λ, μ) is the viscous dissipation rate

$$\int_{\mathcal{O}} \frac{\lambda}{2} (\text{tr}\epsilon[v])^2 + \mu \text{tr}(\epsilon[v]^2) \, dx, \quad \epsilon[v] = \frac{1}{2}(\mathcal{D}v^T + \mathcal{D}v),$$

induced by a velocity field v . This concept yields a natural setting to find paths between shapes, for example, geodesics for the morphing of one shape into another in computer vision. In some cases, the viscous approach is more flexible than the elastic one; in particular, it allows a physically sound modelling of topological changes along a path in shape space as material flowing towards a point from two sides and letting the gap become infinitesimally small. We will employ this concept for the computation of geodesics between shapes in Chapter 6.

While nonlinear elasticity is based on the stored potential energy of reversible deformations, the viscous distance depends on irreversible energy loss during a deformation.

Despite this conceptual difference, a viscous deformation may be seen as the limit of many infinitesimally small elastic deformations with subsequent stress relaxation, where the material stiffness is proportional to the deformation velocity. This connection between both concepts will be of primary importance for the approximation of geodesics in Chapter 6.

The above two shape spaces are highly different from linear vector spaces. Nevertheless, it is sometimes desirable to have some linear representation of a number of shapes, on which for instance a principal component analysis can be performed. In the viscous, Riemannian setting, such a linear representation is provided by the inverse exponential map, which assigns each shape \mathcal{S} a vector S in the tangent space to the shape space at a chosen reference point $\tilde{\mathcal{S}}$. Indeed, the exponential map $\exp_{\tilde{\mathcal{S}}} : \mathcal{T}_{\tilde{\mathcal{S}}}\mathcal{M} \rightarrow \mathcal{M}$, where \mathcal{M} shall denote the Riemannian shape space, maps each $S \in \mathcal{T}_{\tilde{\mathcal{S}}}\mathcal{M}$ onto $\mathcal{S} \in \mathcal{M}$ such that $\tilde{\mathcal{S}}$ and \mathcal{S} are connected by a geodesic with initial velocity S . This mapping is bijective at least in a neighbourhood of $\tilde{\mathcal{S}}$.

In the elastic, non-Riemannian setting, we would like to have an analogue to be able to search for the dominant modes of variation among a given set of shapes. Here, the natural linear representatives of a set of shapes will be their boundary stresses after deformation into a reference shape $\tilde{\mathcal{S}}$, which are elements of the vector bundle on $\tilde{\mathcal{S}}$. This representation will be used in Chapter 5 in order to perform a principal component analysis on given shapes.

1.2 Shape representations

The representation of shapes $\mathcal{S} = \partial\mathcal{O}$ or objects \mathcal{O} just as subsets of \mathbb{R}^d is difficult for numerical treatment in a variational setting [91]. In the simplest case, this would imply a direct tessellation (for example, a triangulation) of the shape interior \mathcal{O} , however, we then face various problems: If different shapes are regarded as deformed configurations of each other, their triangulations should be consistent. Even more problematic, in averaging or morphing problems, the sought shapes are unknown a priori and cannot be triangulated. Furthermore, an irregular triangulation impedes the use of numerical multiscale approaches, an extension of numerical methods designed for shapes to image morphologies becomes complicated, and topology changes of a shape can hardly be handled.

In fact, it is preferable to represent the shapes \mathcal{S} or objects \mathcal{O} by certain functions living on a computational domain $\Omega \subset \mathbb{R}^d$. There are various well-suited alternatives:

- A so-called *single well phase field* is a piecewise smooth function $u : \Omega \rightarrow \mathbb{R}$, which is zero on $\mathcal{S} \cap \Omega$ and close to one everywhere else (see Section 3.2). While the shape itself is a manifold of codimension 1 and zero thickness, the width of its phase field representation is determined by a model parameter ε [8]. In fact, a single well phase field can be used to describe more general free discontinuity problems (see Section 3.2) and thus is also applicable for image morphologies (for which there

is no distinction between foreground and background or interior and exterior of a shape). For this reason, we will employ this shape representation in Chapter 4.

- A *double well phase field* is a smooth function $u : \Omega \rightarrow \mathbb{R}$ which is either close to one or close to minus one and thus discriminates two phases. The width of the interface between regions with $u \approx 1$ and $u \approx -1$ is determined by a scale parameter ε [89] (see Section 3.2). A double well phase field inherently distinguishes between the interior and the exterior of a phase and is therefore well-suited for the representation of shapes $\mathcal{S} = \partial\mathcal{O}$ and the corresponding objects \mathcal{O} . We will use this shape description in the context of shape optimisation in Chapter 7.
- The shape interior can also be described by the zero-super-level set of a *level set function*, which again inherently provides a distinction between interior and exterior of a shape. For numerical effectiveness, the transition from inside to outside is smoothed with the help of a regularised Heaviside function. (Such smoothing can be avoided by the use of shape derivatives combined with a Hamilton–Jacobi equation to describe the evolution of a level set function [21]. However, this prevents the creation of holes inside an object [6].) Unfortunately, unlike for phase fields, the analysis for decreasing regularisation has not been elaborated, yet. However, whereas phase fields more easily align to the spatial discretisation grid (since this is optimal with respect to certain regularisation energies), models based on a level set description of shapes are less prone to artificial anisotropies, which will be of particular importance for the computation of paths of shapes in Chapter 6.

1.3 Coupling of deformations and shape representations

Since shapes are regarded as boundaries of deformable objects, there will naturally arise some kind of coupling between the shape representations (as phase fields or level sets $u : \Omega \rightarrow \mathbb{R}$) and shape deformations $\phi : \Omega \rightarrow \mathbb{R}^d$. Two types of coupling will occur.

First of all, we will encounter constraints that a deformed shape and another one have to match according to $\phi(\mathcal{S}_1) = \mathcal{S}_2$. Such constraints occur during shape averaging or the computation of geodesics, and they imply a concatenation of the shape representation $u : \Omega \rightarrow \mathbb{R}$ with the deformation ϕ . Usually, the constraints can be implemented in a variational problem by adding a mismatch penalty which compares one shape representation u_1 with the pullback $u_2 \circ \phi$ of the other. Consequently, the lower semi-continuity of the model energies and thus existence of minimisers for the variational problems will strongly depend on the regularity of the shape representations and the deformations. The implementation of the constraints as mismatch penalties will allow for an alternating minimisation technique to solve the variational problems: We alternately minimise for the shape representations and for the deformations.

The second type of coupling occurs in the form of PDE constraints for the deformations, which appear in shape optimisation. These PDE constraints can be expressed as an inner variational problem: The deformation ϕ has to be a minimiser of a certain mechanical energy (see Chapter 7), part of which is formed by an elastic deformation energy. Here, the shape representation enters the elastic deformation energy as a multiplicative factor, for example in the case of a double well phase field u as

$$\mathcal{W}[u, \phi] = \int_{\Omega} \left((1 - \delta) \frac{(u + 1)^2}{4} + \delta \right) \hat{W}(\|\mathcal{D}\phi\|_F, \|\text{cof}\mathcal{D}\phi\|_F, \det\mathcal{D}\phi) \, dx.$$

The factor $(1 - \delta) \frac{(u+1)^2}{4} + \delta$ determines the local elastic stiffness: It is roughly one inside the shapes (phase $u \approx 1$) and $\delta \ll 1$ outside ($u \approx -1$). The lower semi-continuity of such energies and thus the existence of minimisers depends crucially on an appropriate, coercive design of the different energy terms (which is here ensured by the regularising δ and the factor being quadratic in u). Such kinds of PDE constraints will require a nested optimisation with sufficient accuracy and robustness requirements for the inner mechanical energy minimisation problem.

2 Related work on shape spaces and shape optimisation

Averages, dominant modes of variation, and paths of shapes or images have been computed in various settings. Applications include the computation of priors for object recognition or segmentation [43, 79, 44], the construction of standardised atlases of the human anatomy [68, 97, 110, 15, 14, 72, 82, 102, 113, 36, 37, 34, 86], pathology detection and mapping of functional to anatomical regions in medical imaging (see for instance [56] for a survey of the potential of shape analysis in brain imaging), shape morphing in computer vision [76], and clustering or classification of shapes [83], just to name a few. Sometimes, the methods were originally developed for the application to images, but can be extended to shapes (apart from some special cases such as joint image registration and image interpolation or blending to simultaneously represent complementary information [72, 74] or the computation of image density preserving mappings for morphing flame or wave images [137]).

There is a broad range of employed models and techniques, and the next sections provide a brief overview over different approaches to model a space of shapes, over variational image registration and segmentation (to which most models for shape analysis are linked via the shape representations as phase fields or level sets), and finally over shape or topology optimisation.

2.1 Different approaches to shape space

The notion of a shape space was introduced by Kendall already in 1984 [75]. He considers shapes as k -tuples of points in \mathbb{R}^d , which can for example be interpreted as discretised curves or nodes of triangulated surfaces. This shape space Σ_d^k is endowed with a quotient metric according to $\Sigma_d^k = \{(\mathbb{R}^d)^k \setminus 0\}/\text{Sim}$, where $(\mathbb{R}^d)^k$ denotes the ordinary (dk) -dimensional Euclidean space and Sim is the group of similarities generated by translation, rotation, and scaling. Kendall furthermore analyses the topology and manifold structure of the space Σ_2^k of discretised planar curves and identifies it (up to a scale change) with the complex projective space $\mathbb{C}P^{k-2}$.

The following paragraphs will describe a list of different shape spaces employed in the literature. We will proceed from finite- and infinite-dimensional normed vector spaces via several nonlinear metric spaces to spaces with a Riemannian structure, which can be further subdivided into finite- and infinite-dimensional approaches. Also, we will

consider manifolds of shapes that are learnt from training shapes.

Often, a shape space is modelled as a linear vector space and is not invariant with respect to shift or rotation a priori (this is sometimes achieved by alignment during preprocessing steps). In the simplest case, such a shape space is made up of vectors of landmark positions, as for example in Kendall’s work. Using the manifold structure described above, Kendall also considers statistical analysis on shapes in $\mathbb{C}\mathbb{P}^{k-2}$. Cootes et al. perform a principal component analysis (PCA) on training shapes with consistently placed landmarks to obtain priors for edge-based image segmentation [43]. Hafner et al. use a PCA of position vectors covering the proximal tibia to reconstruct the tibia surface just from six dominant modes [68]. Perperidis et al. automatically assign consistent landmarks to training shapes by a non-rigid registration as a preprocessing step for a PCA of the cardiac anatomy [97]. Söhn et al. compute dominant eigenmodes of landmark displacement on human organs, also using registration for preprocessing [110].

As an infinite-dimensional vector space, the Lebesgue-space L^2 has served as shape space, where again shape alignment is a necessary preprocessing step. Leventon et al. identify shapes with their signed distance functions and impose the Hilbert space structure of L^2 on them to compute an average (which in general is no signed distance function) and dominant modes of variation [79]. They only consider the planar case to obtain shape priors for geodesic active contours image segmentation. Tsai et al. apply the same technique to 3D prostate images [119]. Dambreville et al. also compute shape priors, but using characteristic instead of signed distance functions [47].

A more sophisticated, but still not rigid body motion invariant shape space is obtained by considering shapes as subsets of a metric space (\mathbb{R}^d , for example), endowed with the Hausdorff distance

$$d_H(\mathcal{S}_1, \mathcal{S}_2) = \max\left\{\sup_{x \in \mathcal{S}_1} \inf_{y \in \mathcal{S}_2} d(x, y), \sup_{y \in \mathcal{S}_1} \inf_{x \in \mathcal{S}_2} d(x, y)\right\}$$

between any two shapes $\mathcal{S}_1, \mathcal{S}_2$, where $d(\cdot, \cdot)$ denotes the metric of the ambient space. Charpiat et al. employ smooth approximations of the Hausdorff distance based on a comparison of the signed distance functions of shapes [31]. They investigate the correlation of different shapes via gradient descent type morphing from one shape onto the other. The gradient of the shape distance functional is decomposed into rigid body motions, scalings, and the remainder. A separate weighting of the different components mimics frame indifference during gradient descent warping. For a given set of shapes, the gradient at the average shape is regarded as shape variation of the average and used to analyse dominant modes in the variation of the averaged shape [29].

An isometrically invariant distance measure between shapes (or more general metric spaces) is provided by the Gromov–Hausdorff distance, which can be defined as

$$d_{GH}(\mathcal{S}_1, \mathcal{S}_2) = \frac{1}{2} \inf_{\phi: \mathcal{S}_1 \rightarrow \mathcal{S}_2} \sup_{\psi: \mathcal{S}_2 \rightarrow \mathcal{S}_1} \sup_{\substack{y_i = \phi(x_i) \\ \psi(y_i) = x_i}} |d_{\mathcal{S}_1}(x_1, x_2) - d_{\mathcal{S}_2}(y_1, y_2)|,$$

where $d_{\mathcal{S}_i}(\cdot, \cdot)$ is a distance measure between points in \mathcal{S}_i . The Gromov–Hausdorff distance represents a global, supremum-type measure of the lack of isometry between two shapes, which makes it difficult to locate or locally examine isometry distortions. Memoli and Sapiro use this distance for clustering shapes described by point clouds, and they discuss efficient numerical algorithms to compute Gromov–Hausdorff distances based on a robust notion of intrinsic distances $d_{\mathcal{S}}(\cdot, \cdot)$ on the shapes [83]. Bronstein et al. incorporate the Gromov–Hausdorff distance concept in various classification and modelling approaches in geometry processing [20]. Memoli investigates the relation between the Gromov–Hausdorff distance and the Hausdorff distance under action of Euclidean isometries as well as L^p -type variants of the Gromov–Hausdorff distance [84].

A shape space can also have the structure of a Riemannian manifold. In this case, paths, path lengths, and shortest paths (geodesics) are generically defined. Averaging of given points \mathcal{S}_i , $i = 1, \dots, n$, on the manifold can be performed via the generalisation of the (geometric) mean

$$\mathcal{S} = \arg \min_{\tilde{\mathcal{S}}} \sum_{i=1}^n d(\tilde{\mathcal{S}}, \mathcal{S}_i)^2,$$

which is due to Fréchet [60] and was further analysed by Karcher [73]. Here, $d(\cdot, \cdot)$ represents the geodesic distance, and the mean will satisfy $\sum_{i=1}^n \log_{\mathcal{S}}(\mathcal{S}_i) = 0$. Fletcher et al. propose to use the more robust shape median $\arg \min_{\mathcal{S}} \sum_{i=1}^n d(\mathcal{S}, \mathcal{S}_i)$ instead of the geometric mean and compute it numerically by a step size-controlled gradient descent in the case of planar curves [59]. In a Riemannian shape space, there is also a natural linear representation of shapes in the tangent space at the Fréchet mean via the log-map, which enables a PCA.

One particular Riemannian shape space is given by the space of polygonal medial axis representations, where each shape is described by a polygonal lattice and spheres around each vertex [135]. Of course, the lattice topologies of different shapes has to be consistent. Fletcher et al. exploit the Lie group structure of the medial representation space to approximate the Fréchet mean as exponential map of the average of the logarithmic maps of the input, and they perform a PCA on these log-maps to obtain the dominant geometric variations of kidney shapes [57] and brain ventricles [58]. However, they do not quotient out rigid body motions. Fuchs and Scherzer use the PCA on log-maps to obtain the covariance Σ of medial representations composed of just two atoms. They define a Mahalanobis distance via this covariance to obtain priors for edge based image segmentation [63, 61]. This Mahalanobis distance imposes a new metric on the shape manifold: An orthonormal basis of the tangent space at the mean is transported along geodesics, providing a canonical isometry between tangent spaces at different points. In each tangent space, the new metric is induced by the inner product $g(v, w) = v^T \Sigma^{-1} w$ for tangent vectors v, w in the transported basis representation.

Riemannian shape spaces have also been devised for triangulated surfaces. As in landmark space, a consistent triangulation of different shapes is essential. Kilian et al. compute and extrapolate geodesics between triangulated surfaces using isometry invari-

ant Riemannian metrics (on vector fields on vertices) that measure the local distortion of the grid (or rather of its intrinsic metric) [76]. Eckstein et al. employ different metrics in combination with a smooth approximation to the Hausdorff distance to perform gradient flows for shape matching [53].

An infinite-dimensional Riemannian shape space has been developed for planar curves. Klassen et al. propose a framework for geodesics in the space of arclength parameterised curves and implement a shooting method to find them [77]. As Riemannian metric, they use the L^2 -metric on variations of the direction or curvature functions of the curves. Schmidt and Cremers present an alternative variational approach for the computation of these geodesics [107]. Srivastava et al. assign different weights to the L^2 -metric on stretching variations and bending variations and obtain an elastic model of curves [112]. Michor and Mumford examine Riemannian metrics on the manifold of smooth regular curves [85], expressed as the orbit space of C^∞ -embeddings of S^1 into \mathbb{R}^2 under the action of diffeomorphic reparameterisation. They show the L^2 -metric in tangent space to be pathologic in the sense that it leads to arbitrarily short geodesic paths, a particular instance of the failure of the Hopf–Rinow theorem in infinite dimensions. They hence employ a curvature-weighted L^2 -metric instead. For the same reason, Sundaramoorthi et al. use Sobolev metrics in the tangent space of planar curves to perform gradient flows for image segmentation via active contours [114]. Finally, Younes considers a left-invariant Riemannian distance between planar curves by identifying shapes with elements of a Lie group acting on one reference shape [133].

As already mentioned in the previous chapter, a not purely geometric view of shapes as curves or surfaces is to consider them as boundaries $\partial\mathcal{O}$ of objects $\mathcal{O} \subset \mathbb{R}^d$ which are all connected by homeomorphisms. A corresponding Riemannian structure is then obtained by identifying the tangent space at $\mathcal{S} = \partial\mathcal{O}$ with velocity fields $v : \mathcal{O} \rightarrow \mathbb{R}^d$ and defining a metric on these. Maps between different shapes are then generated via integration of velocity fields along geodesics. For sufficient regularity of the Riemannian metric $g(\cdot, \cdot)$, these maps are actually diffeomorphisms, for example, if $g(v, v) = \int_{\Omega} Lv \cdot v dx$ for a higher order elliptic operator L . For the numerical computation of path lengths, usually the velocity field is discretised in time, yielding rigid body motion invariance (if the Riemannian metric has that property) and 1-1 correspondence between shapes only in the limit for infinitesimal time step size. Dupuis et al. exploit the fact that for sufficient Sobolev regularity of the motion field, the induced flow consists of a family of diffeomorphisms that can be used for image matching in a shooting approach fashion [52]. Beg et al. examine the corresponding Euler–Lagrange equations for the velocity field in this shooting problem and implement a gradient descent algorithm for it [13]. Miller and Younes consider the space of registered images as the product space of the Lie group of diffeomorphisms and images and define a Riemannian metric using sufficiently regular elliptic operators on the generating velocity fields, which may also depend on

the current image [87, 88]. Fuchs et al. propose a viscous Riemannian metric

$$g(v, v) = \int_{\mathcal{O}} \frac{\lambda}{2} (\operatorname{tr} \epsilon[v])^2 + \mu \operatorname{tr}(\epsilon[v]^2) dx, \quad \epsilon[v] = \frac{1}{2}(\mathcal{D}v^T + \mathcal{D}v),$$

just on the object (and not the ambient space), which measures the infinitesimal change of area weighted by λ and length weighted by μ on objects \mathcal{O} . They compute geodesics between two planar shapes based on a triangulation of one of them [62].

As already explained earlier, we will employ the same Riemannian concept in Chapter 6. However, Chapters 4 and 5 are based on a different, nonlinearly elastic concept, where we measure similarity between shapes via a nonlinear elastic deformation energy. A related approach is due to Hong et al. [71] who use the stored energy from linearised elasticity as a distance measure between shapes in order to compute shape averages and principal modes of shape variation. Their energy functional is invariant with respect to rigid body motions only in an infinitesimal sense, and it measures deformations not only within the objects \mathcal{O} , but rather on the whole ambient space. Pennec et al. [95, 94] define a nonlinear elastic energy as the integral over the ambient space of a St. Venant–Kirchhoff-type energy density that depends on the logarithm of the Cauchy–Green strain tensor $\mathcal{D}\phi^T \mathcal{D}\phi$ for a deformation ϕ . This energy induces a symmetric distance between diffeomorphisms and can be used as the regularising prior in nonlinear registration of images. Furthermore, it also measures isometry violation and acts as a penalty to avoid material interpenetration. However, it is in general not quasi-convex, which renders the existence theory of minimisers difficult.

Finally, a shape space is sometimes understood as a manifold, learnt from training shapes and embedded in a higher-dimensional (often linear) space. In the general setting, the embedding space is typically not specified: The approach is often applied to grey value images, but application to landmarks or curves, for example, is also possible. The manifold is learnt from training shapes or images and is usually used as a prior in image segmentation. Chalmond and Girard approximate a point cloud in a higher-dimensional space by a d -dimensional manifold, defined as the tensor product of d B-splines [23]. They are thus able to represent also truly nonlinear transformations of learnt 2D images of 3D objects. Many related approaches are based on kernel density estimation in feature space, where the manifold is described by a probability distribution in the embedding space. This probability distribution is computed by mapping points of the embedding space into a higher-dimensional feature space and assuming a Gaussian distribution there. This map into feature space is usually unknown, and points in feature space (for example, the mean of the feature space representations of the training shapes) in general have no preimage in shape space so that approximate preimages have to be obtained via a variational formulation [100]. Cremers et al. use this technique to obtain 2D silhouettes of 3D objects as priors for image segmentation [44]. Rathi et al. provide a comparison between kernel PCA, local linear embedding (LLE, that is, approximation of the manifold near a point as convex hull of nearest neighbour training shapes), and kernel LLE (kernel PCA only on the nearest neighbours) [99]. Thorstensen et al. approximate

the shape manifold using weighted Karcher means of nearest neighbour shapes obtained by diffusion maps [117].

2.2 Registration

The variational problems treated in the following chapters involve some kind of registration, that is, computation of matching deformations between shapes or their phase field or level set representations. Registration is always based on a particular choice of a similarity measure and a deformation regularisation, which prevents arbitrarily irregular deformations.

The similarity measures employed in the literature range from landmark-based measures [35, 43] and image intensity-based measures [34, 69, 70, 74] over joint entropy or mutual information [14, 101, 113, 121, 126, 129] (see [55, 54] for an analytical discussion and comparison of these methods) to morphology-based similarity measures [50, 51].

The regularisation of the matching deformation is most easily achieved by restricting its degrees of freedom, for example, by allowing only rigid deformations [126, 132], affine deformations [129], deformations described by B-splines [14, 111], or clamped-plate splines [82]. A different way is to employ a variational regularisation, that is, an energy which tends to infinity for irregular deformations (for an overview over the employed methods in medical image registration, see [116, 118]). As regularisation energies one can use the path length in the space of diffeomorphisms with a sufficiently regular Riemannian metric [87, 88], linearised elasticity [37, 74, 86], nonlinear hyperelasticity [50, 51, 98], or viscous fluid regularisation [19, 36, 35, 34]. As explained in the previous chapter, the latter two regularisations are particularly related to the methods employed in this thesis.

Computing an average of a given set of shapes (Chapter 4) is in particular related to the simultaneous registration of a number of shapes or images to one single reference object, which has also been attempted in various settings [14, 15, 82, 132, 113].

2.3 Segmentation

Shapes are frequently encoded in images or volume data. Computations in shape space are thus often closely linked to segmentation, that is, the partitioning of an image into different regions. There is an enormous body of literature based on the seminal paper by Mumford and Shah [92], in which a model is proposed to extract a cartoon from an image via a variational free discontinuity problem. For a given image $y^0 : \Omega \rightarrow \mathbb{R}$, $\Omega \subset \mathbb{R}^d$, they propose to minimise the energy

$$\mathcal{E}_{\text{MS}}[y, K] = \int_{\Omega \setminus K} |\nabla y|^2 + \alpha |y - y^0|^2 dx + \nu \mathcal{H}^{d-1}(K)$$

to obtain a piecewise smooth approximation y and an image edge set K . Here, \mathcal{H}^{d-1} shall denote the $(d - 1)$ -dimensional Hausdorff measure. The theoretic examination of this energy with existence results is quite elaborate [90, 49, 46], and a number of numerically tractable model approximations have been formulated. A very successful approach due to Ambrosio and Tortorelli encodes the edge set K as a single well phase field [7] (a detailed explanation is given in Section 3.2). We will employ this method in conjunction with shape averaging in Chapter 4. An alternative approximation, which segments the image into (possibly multiple) disjoint regions, is based on a Modica–Mortola-type double well phase field [89] (see Section 3.2). Another widely used approximation to the Mumford–Shah energy is due to Chan and Vese and uses level set functions $u : \Omega \rightarrow \mathbb{R}$ to distinguish between different image regions. It was originally formulated for image segmentation into two regions with constant grey values c_1, c_2 by minimising

$$\mathcal{E}_{CV}[c_1, c_2, u] = \int_{\Omega} H_{\varepsilon}(u)|y^0 - c_1|^2 + (1 - H_{\varepsilon}(u))|y^0 - c_2|^2 dx + \nu \int_{\Omega} |\nabla H_{\varepsilon}(u)| dx,$$

where H_{ε} is a regularised Heaviside function and the last term approximates the $(d - 1)$ -dimensional Hausdorff measure \mathcal{H}^{d-1} of the zero-level set of u [25, 27, 28]. Meanwhile, it has been extended to allow for more than two regions and piecewise smooth (instead of piecewise constant) image approximations [26]. We will borrow this idea of describing objects as concatenations of a regularised Heaviside function with level set functions for the computation of geodesics in Chapter 6.

It has been observed that image segmentation and image registration can benefit from each other if performed simultaneously: The quality of a registration of segmented images of course strongly depends on the robustness of the segmentation, while a registration can also help to improve the segmentation result of an image due to complementary information from another image (for example, if edge information in one image is destroyed by noise). This effect has been exploited in joint segmentation and registration approaches in [129], in the context of geodesic active contours using level sets [132, 120], and for the registration of image morphologies [51]. Similarly, Young and Levy use the segmentation of previous images to guide the edge detection in consecutive images [134]. We will also make use of this effect during the computation of shape averages in Chapter 4.

2.4 Shape optimisation

Chapter 7 will complement the shape space analysis from Chapters 4 to 6 in the direction of classical shape optimisation. Here, we will adopt the more general view of shape averaging or computation of shape geodesics as basically being optimisation tasks to be performed in shape space. These optimisation problems are special in that their objective functionals will include the deformation energies of optimal deformations between different shapes (since these deformation energies are the ones imposing the elastic

or viscous structure on shape space). This fact automatically ensures that the shape deformations which result from the optimisation are physically meaningful as they are minimisers of the mechanical energy. However, for more general optimisation problems in which shapes are still to be interpreted as boundaries of deformable objects, we will have to impose additional constraints on the shape deformations that ensure their physical meaningfulness. Classical shape optimisation can be regarded as the most classical such problem and is therefore treated in Chapter 7. This section is intended to introduce the corresponding related work.

Shape optimisation typically aims at finding an open domain $\mathcal{O} \subset \mathbb{R}^d$ within a set of admissible domains that minimises a functional which is usually of the form

$$\mathcal{J}[v[\mathcal{O}], \mathcal{O}] = \int_{\Omega} f(v[\mathcal{O}], \chi_{\mathcal{O}}) dx$$

for some f , where $\Omega \subset \mathbb{R}^d$ is the domain of interest, $\chi_{\mathcal{O}}$ denotes the characteristic function of \mathcal{O} , and $v[\mathcal{O}]$ solves some partial differential equation on \mathcal{O} . $v[\mathcal{O}]$ can represent a temperature distribution or an elastic displacement, for example. Most often, not only the position and geometry of the shape contour $\partial\mathcal{O}$, but also the topology of \mathcal{O} is subject to optimisation.

For the optimisation of elastic structures, $v[\mathcal{O}]$ typically solves the equations of linearised elasticity,

$$\begin{aligned} \operatorname{div} \sigma &= 0 && \text{in } \mathcal{O}, \\ v &= 0 && \text{on } \Gamma_D, \\ \sigma \nu &= F && \text{on } \Gamma_N, \\ \sigma \nu &= 0 && \text{on } \partial\mathcal{O} \setminus (\Gamma_D \cup \Gamma_N), \end{aligned}$$

where the Cauchy stress σ is given by $\mathbf{C}\epsilon[v] = \frac{1}{2}\mathbf{C}(\mathcal{D}v + \mathcal{D}v^T)$ for the fourth-order elasticity tensor \mathbf{C} , F represents some surface loading, and $\Gamma_D, \Gamma_N \subset \partial\mathcal{O}$ are fixed parts of the boundary, whose normal is given by ν (we ignore volume forces for simplicity). The range of objective functionals $\mathcal{J}[v, \mathcal{O}]$ is relatively diverse. The mechanical work of the load, the so-called compliance $\frac{1}{2} \int_{\Gamma_N} F \cdot v da$, is very popular [3, 4, 5, 2, 123, 122, 67, 125, 124] since it equals the energy to be absorbed by the elastic structure. A related choice is the L^2 -norm of the internal stresses [4, 5, 2], $\int_{\mathcal{O}} \|\sigma\|_F^2 dx$. If a specific displacement v_0 is to be reproduced, then the L^2 -distance $\int_{\Omega} |v - v_0|^2 dx$ serves as the appropriate objective functional [5]. Other possibilities include functionals depending on the shape eigenfrequencies or the compliance for design-dependent loads [6, 17, 109]. If the elastic structure has to bear multiple loads, then $\mathcal{J}[v, \mathcal{O}]$ can be taken as a weighted sum of the structural responses (such as the compliances) under the different loads. For example, Conti et al. minimise the expected value of the compliance [41]. They assume the different loads to be linear combinations of a set of basis loads so that—due to the linearity of the equations—the structural response to these basis loads suffices for the

computation of the expected value. Of course, the weighted sum can also be replaced by some nonlinear relation: An excess probability or expected excess for the compliance is minimised in [42] to obtain risk-averse shape designs. As a final example, retrieving the equilibrium microstructure of a compound material (such as a binary alloy) can also be regarded as an optimisation of elastic shapes: Here, $\mathcal{J}[v, \mathcal{O}]$ represents the chemical bulk energy and the interfacial energy of the multiphase material as well as the elastic deformation energy due to the lattice misfit between the different phases. Garcke [64, 65] has examined the evolution of this energy via a Cahn–Hilliard model for Ostwald ripening and also proved existence and uniqueness in the case of linearised elasticity.

Typically, the optimisation problem is complemented by a volume constraint for \mathcal{O} (otherwise, especially for compliance minimisation, $\mathcal{O} = \mathbb{R}^d$ would be optimal). An equality constraint $|\mathcal{O}| = V$ is either ensured by a Cahn–Hilliard-type H^{-1} -gradient flow [64, 65, 136] or a Lagrange multiplier ansatz [6, 17, 80]. A quadratic penalty term or an augmented Lagrange method is employed in [125]. An inequality constraint $|\mathcal{O}| \leq V$ is implemented in [123, 122, 67], using a Lagrange multiplier. Chambolle [24] exploits the monotonicity of the compliance \mathcal{C} (in sense $\mathcal{C}(\mathcal{O}_1) \geq \mathcal{C}(\mathcal{O}_2)$ for $\mathcal{O}_1 \subset \mathcal{O}_2$) to replace the equality by an inequality constraint. Finally, the volume may just be added to the objective functional as an additional cost $\nu|\mathcal{O}|$ for some parameter ν [2]. Allaire et al. [3] here interpret ν as the Lagrange multiplier for a volume equality constraint, hoping that it can be tuned to obtain the desired volume (however, they cannot establish continuity, but only monotonicity of the dependence of $|\mathcal{O}|$ on ν).

The above shape optimisation problems are generically ill-posed since microstructures tend to form [3], which are associated with a weak but not strong convergence of the characteristic functions $\chi_{\mathcal{O}_i}$ along a minimising sequence \mathcal{O}_i . In particular, rank- d sequential laminates with the lamination directions aligned with the stress eigendirections are known to be optimal for compliance minimisation, but there are also other optimal microstructures such as the so-called concentric spheres construction under hydrostatic pressure [3]. To complicate matters even further, the 2D optimum microstructure generally differs from the optimal plane stress configuration in 3D, and in the case of multiple loads the optimum is only achieved for rank-3 sequential laminates in 2D and rank-6 sequential laminates in 3D (just as needed to reproduce isotropic material properties) [3].

The above ill-posedness calls for regularisation, for which there are several possibilities. In the simplest case, the spatial discretisation yields a well-posed problem, however, the optimisation results will be highly mesh-dependent. As a more adequate possibility, in 2D one can restrict the number of holes in \mathcal{O} : Chambolle [24] shows that on two-dimensional bounded open sets \mathcal{O} whose complement has a finite number of connected components, $[H^1(\mathcal{O})]^2$ is dense in vector fields with symmetrised gradient in $L^2(\mathcal{O})$, and he uses this to show existence of a solution to the compliance minimisation problem. A different approach is to penalise the shape perimeter by adding a term $\eta\mathcal{H}^{d-1}(\partial\mathcal{O})$ to the objective functional, which (if the void is replaced by some weak material) also results in existence of optimal shapes as studied, for example, in [9] for a scalar problem.

They exploit the compactness property that for any sequence of sets \mathcal{O}_i with bounded perimeter, the characteristic functions of a subsequence converge strongly in $L^1_{\text{loc}}(\mathbb{R}^d)$. The optimum shape for both types of regularisation is quite likely to be the same in most cases, although the optimum \mathcal{O} would be allowed to possess an infinite sequence of holes of decreasing size for a perimeter penalisation, while for a restricted number of holes, these holes might have arbitrarily irregular boundaries. A further method consists in the problem relaxation: The set of admissible shapes can be extended to allow for microstructures, and a quasiconvexification of the integrand in $\mathcal{J}[v, \mathcal{O}]$ (by taking the infimum over all possible microstructures) then also ensures existence of minimisers [3]. An alternative is given by the integrand convexification, where instead of material with microstructures we allow for a material with an intermediate density u between full material and void whose elasticity tensor is given by $u^p \mathbf{C}$ for some $p \leq 1$.

There are various approximations and implementations of the elastic shape optimisation problem, each of which more or less corresponds to a particular type of regularisation. A direct triangulation of \mathcal{O} or its boundary would probably work with all regularisations, but it requires remeshing during the optimisation and induces technical difficulties with topological changes. As a related, more feasible variant, Conti et al. employ a level set description of \mathcal{O} and compute the displacement on \mathcal{O} using composite finite elements [41, 42].

The so-called evolutionary structural optimisation (ESO) is based on discretising the computational domain by finite elements and successively removing those elements which contribute least to the structural stiffness (or another chosen objective, see for example [10]). This corresponds to a regularisation via discretisation and thereby introduces a mesh-dependence.

In level set approaches, the set \mathcal{O} is described as the zero-super-level set of a level set function $u : \Omega \rightarrow \mathbb{R}$. The optimisation of u then is usually performed by some descent algorithm based on the shape derivative $\delta_{\mathcal{O}} \mathcal{J}[v[\mathcal{O}], \mathcal{O}]$, which is defined as an operator on velocity fields $w : \Omega \rightarrow \mathbb{R}^d$ via

$$\langle \delta_{\mathcal{O}} \mathcal{J}[v[\mathcal{O}], \mathcal{O}], w \rangle = \left. \frac{dJ[v[(\text{id} + \delta w)(\mathcal{O})], (\text{id} + \delta w)(\mathcal{O})]}{d\delta} \right|_{\delta=0}.$$

Actually, the velocity w here only has to be defined on the boundary $\partial\mathcal{O}$. Burger [21] proposes a gradient flow framework to obtain velocities w of the boundary for which the energy decreases: w is chosen such that

$$\langle \delta_{\mathcal{O}} \mathcal{J}[v[\mathcal{O}], \mathcal{O}], \theta \rangle = g(w, \theta)$$

for some inner product g and all test velocities θ . Burger also provides examples of the resulting flow for different g such as the inner product in $H^1(\partial\mathcal{O})$ (Laplace–Beltrami flow), $H^{1/2}(\partial\mathcal{O})$ (Stefan flow), $L^2(\partial\mathcal{O})$ (Hadamard flow), $H^{-1/2}(\partial\mathcal{O})$ (Mullins–Sekerka flow), and $H^{-1}(\partial\mathcal{O})$ (surface diffusion flow). Allaire et al. [6] also propose to use H^1 ,

L^2 -, or H^{-1} -type inner products. The boundary velocity w is then translated into an update of the level set function u via the Hamilton–Jacobi equation

$$\frac{\partial u}{\partial t} + w \cdot \nabla u = 0,$$

where $w \cdot \nabla u$ may also be replaced by $w_n |\nabla u|$ for the normal velocity w_n of the boundary. For this purpose, the velocity field w is usually extended from $\partial\mathcal{O}$ onto Ω , and the Hamilton–Jacobi equation is solved in a narrow band around $\partial\mathcal{O}$ by an explicit upwind scheme [2, 123] (Liu et al. [80] additionally introduce a regularising diffusion term into the Hamilton–Jacobi equation). In order to circumvent the CFL-restriction of the time step, Xia et al. introduce a semi-Lagrange scheme which is basically an upwind scheme along the characteristics of the Hamilton–Jacobi equation [131]. In two dimensions, this level set approach corresponds to a regularisation by restricting the number of holes in \mathcal{O} : Indeed, during the gradient flow, topology changes can only happen by merging or eliminating holes (whereas in 3D, holes may appear by pinching a thin wall) [2, 6] so that the maximum number of holes is prescribed by the initialisation. This is certainly the reason why Allaire et al. [5, 6] do not need any additional regularisation for the minimum compliance design of a cantilever or the design of a gripping mechanism. Wang et al. [123] also use this level set technique in two and three dimensions and prove that the shape sequence generated by a gradient descent-type update is indeed descending (note, however, that the problem of existence of a minimiser without any further regularisation is still open in 3D). Allaire et al. [6] extend the method for design-dependent surface loads by approximating these as the product of some volume forces and a smoothed Dirac function centred around $\partial\mathcal{O}$. Furthermore, they compute one example based on the geometrically nonlinear St. Venant–Kirchhoff material law.

A different approach (which we will also employ in Chapter 7) is to describe the shapes by a so-called phase field u , whose origin lies in the physical description of multiphase materials: The chemical bulk energy of the material is given by $\int_{\Omega} \Psi(u) dx$ for some potential Ψ with minima $\Psi(u) = 0$ at $u = -1$ and $u = 1$, representing two material phases. This energy is perturbed by an interfacial energy of the form $\int_{\Omega} |\nabla u|^2 dx$. A weighting of both terms according to

$$\frac{1}{2} \int_{\Omega} \varepsilon |\nabla u|^2 + \frac{1}{\varepsilon} \Psi(u) dx$$

induces an energy minimising, optimal profile of u perpendicular to the interface and a transition region whose width scales with the parameter ε (details will be given in Section 3.2). For $\varepsilon \rightarrow 0$, the above integral forces the phase field u towards the pure phases -1 and 1 , and it Γ -converges against the total interface length. For this reason, the approach lends itself for a perimeter regularisation: The above approximation of the perimeter is simply added to the objective functional \mathcal{J} . This technique is employed by Wang and Zhou [124] who minimise the compliance of an elastic structure using a triphasic phase field (with one void and two material phases) for which the potential Ψ

is equipped with a periodically repeated sequence of three minima to allow for all three possible types of phase transitions. Furthermore, they replace the term $\int_{\Omega} |\nabla u|^2 dx$ by an edge-preserving smoothing and perform a multiscale relaxation, starting with large ε and successively decreasing it (however, they seem to decrease ε beyond the point up to which the grid can still resolve the interface). Zhou and Wang [136] compute the Cahn–Hilliard evolution of the shape to be optimised, also using a multiphase material. They solve the elastic equations with finite elements and the resulting fourth-order Cahn–Hilliard-type partial differential equation with a Crank–Nicolson finite difference scheme in which nonlinear terms are approximated by Taylor series expansion and the resulting linear system is solved by a multigrid V-cycle. Burger and Stainko [22] minimise the volume $|\mathcal{O}|$ under a stress constraint and show existence of a corresponding minimiser. They use a double obstacle potential Ψ to reformulate the shape optimisation as a quadratic programming problem with linear constraints. Finally, Bourdin and Chambolle [17] find minimum compliance designs for (design-dependent) pressure loads, using a solid, liquid, and void phase, which they describe by a scalar phase field allowing for the transitions void-solid-liquid. They also prove existence of minimisers for the sharp interface model and implement the optimisation as a semi-implicit descent scheme with linear finite elements on a triangular unstructured mesh.

The so-called homogenisation method corresponds to the relaxation of the optimisation problem. Here, a function u describes the (relative) material density (analogous to the phase field), but it is not forced to take either the value 0 or 1. Instead, a microperforated material of intermediate density is explicitly allowed, whose material properties can (for compliance minimisation) be computed explicitly based on homogenisation and the fact that the optimal microstructures are sequential laminates (whose material properties are known [3]). The void is usually replaced by a very soft “ersatz material” as thoroughly justified in [3]. Allaire et al. [3] express the compliance minimisation problem in terms of macroscopic stresses σ as

$$\min_{\sigma \in \Sigma(\Omega)} \int_{\Omega} \min_{0 \leq u \leq 1} \left(\min_{\mathbf{C} \in L_u} \mathbf{C}^{-1} \sigma : \sigma + \nu u \right) dx$$

with L_u the set of elasticity tensors of sequential laminates with density u and $\Sigma(\Omega) = \{\sigma \in L^2(\Omega) : \operatorname{div} \sigma = 0 \text{ in } \Omega, \sigma \nu = F \text{ on } \Gamma_N\}$. They optimise u by alternately computing σ for a fixed density u and elasticity tensor \mathbf{C} (using finite elements) and then computing the optimal u and \mathbf{C} explicitly for fixed σ . Arising checkerboard instabilities are alleviated by an averaging over adjacent elements. Allaire et al. [4] compute a design which minimises the L^2 -norm of the internal stresses, for which case the optimal microstructures are unknown, unfortunately. They use so-called corrector tensors P_{ε} to describe the material microstructure such that $\sigma_{\varepsilon} - P_{\varepsilon} \sigma \rightarrow_{\varepsilon \rightarrow 0} 0$ in $L^2(\Omega)$, where ε is the scale of the microstructure and σ_{ε} and σ are the corresponding microscopic and macroscopic stress tensors. The L^2 -norm of the microscopic stress can then be described in the limit as $\int_{\Omega} \|P\sigma\|_F^2 dx$, where the stress amplification factor P^2 represents the weak limit of P_{ε}^2 in the case of sequentially laminated microstructures (for other microstructures,

the weak limit is unknown so that the results may be suboptimal). For optimisation, Allaire et al. discretise the set of all possible sequential lamination directions and perform a gradient descent for the density u and the lamination parameters m_i belonging to the different discretised lamination directions. In fact, they minimise $\int_{\Omega} \|P[u, m_i]\sigma\|_F^2 dx$ where $P[u, m_i]$ can be computed explicitly. The advantage of the homogenisation method is that it allows to find a global optimum [2], however, a microperforated structure \mathcal{O} was not intended originally. For this reason, the optimisation is often followed by a (purely numerical and mesh-dependent) postprocessing step in which composite regions are penalised. Allaire et al. [3], for example, do some final iterations replacing the actually optimal density u by $\frac{1}{2}(1 - \cos(u\pi))$.

Another type of methods is subsumed under the term fictitious material approach. Here, instead of computing the elasticity tensor of the optimal laminate, a material density u is associated with the elasticity tensor $u^p \mathbf{C}$ of a fictitious material, which for $p \leq 1$ corresponds to a convexification of the integrand of \mathcal{J} [3]. This procedure seems less adapted than the homogenisation method since it only yields an optimal density, but not an optimal microstructure. Also, a postprocessing with penalisation of intermediate densities is still necessary. For this reason, p is sometimes chosen larger than one (SIMP method, solid isotropic material with penalization, see for example [124, 22, 123]) so as to prefer either stiff material with $u \approx 1$ or the density $u \approx 0$ with no volume costs. In fact, the fictitious material approach is closely related to the phase field method, since the phase field u can also be interpreted as a generalised material density. Also, in phase field methods, intermediate values of u are also assigned a fictitious elasticity tensor of the above type (see Chapter 7), however, the significance of this tensor vanishes as u is forced to the pure phases.

Finally, a number of articles put forward improvements to or mixtures of the above methods. For example, in order to accelerate convergence of the shape topology or to allow the level set method to create holes in 2D, the topological derivative is sometimes used to identify and remove rather inactive interior material parts [42, 67]. Wang et al. [122] compute minimum compliance designs applying the SIMP method. For regularisation, they add an isotropic diffusion term $\int_{\Omega} \varphi(\nabla u) dx$ to the objective functional (for example, $\varphi(\nabla u) = |\nabla u|^2$) which almost yields a phase field type model (only the chemical potential is missing). They compare various nonlinear and nonconvex diffusion laws to preserve edges and implement an explicit gradient flow as well as a fixed point iteration in u , applying the diffusion after each iteration. Similarly, instead of smoothing u , Guo et al. [67] take the stiffness on an element as the average over its nodes. They describe the characteristic function of \mathcal{O} by the concatenation of a smoothed Heaviside function with a level set function, where the smoothed Heaviside function acts like a phase field profile. Wei and Wang [125] encode \mathcal{O} by a piecewise constant level set function u , which is also closely related to the phase field method: They regularise u via total variation, which in conjunction with the penalty $\int_{\Omega} (u-1)^2(u-2)^2 dx$ for the constraint $u \in \{1, 2\}$ has a similar effect as the phase field perimeter term $\frac{1}{2} \int_{\Omega} \varepsilon |\nabla u|^2 + \frac{1}{\varepsilon} \Psi(u) dx$. Xia and Wang [130] compute functionally graded structures, where the shape is de-

scribed by a level set function and the smoothly varying material properties by a scalar field (that actually describes the mixture of two material components from which the physical properties are computed under an isotropy assumption). Abe et al. [1] use a boundary element method instead of finite elements to solve the elastic equations.

3 Concepts from elasticity and multiphase modelling

Throughout this work, our notion of shapes will be based on their interpretation as boundaries of physically deformable objects. A physically sound modelling of these deformations will play an essential role, for which reason the following section introduces the basic ideas of solid continuum mechanics (for a detailed introduction into the topic we refer to [81, 38, 93]). Furthermore, the description of shapes via phase fields, which we will use in Chapters 4 and 7, also deserves some introductory notes, which are given in Section 3.2.

3.1 Solid continuum mechanics

Consider a body of solid material, represented by an open, bounded, connected subset $\mathcal{O} \subset \mathbb{R}^d$ with Lipschitz-boundary, where we restrict to $d = 3$ (the two-dimensional case can be treated analogously). Assume the body to be fixed at part of its boundary, $\Gamma_D \subset \partial\mathcal{O}$, and to be subjected to a surface load F at $\Gamma_N \subset \partial\mathcal{O}$, where $\Gamma_D \cap \Gamma_N = \emptyset$ (we shall neglect body forces for simplicity). The surface load $F : \Gamma_N \rightarrow \mathbb{R}^d$ has an interpretation of force per surface area, and it induces a deformation $\phi : \mathcal{O} \rightarrow \mathbb{R}^d$ of the body such that each point $x \in \mathcal{O}$ is displaced to $\phi(x)$ (Figure 3.1). Our aim is to describe this deformation and the energy associated with it.

We postulate the existence of a Gibbs free energy density $W[\phi]$ of the deformed material, which only depends on the position $x \in \mathcal{O}$ and on the Jacobian $\mathcal{D}\phi$ of the deformation, the so-called deformation gradient. Materials for which this assumption holds are called hyperelastic. The frame indifference principle requires that the local elastic energy is independent of the frame of reference, that is, the underlying coordinate system. Hence, any coordinate transform $y = Qx + b$ for a rotation $Q \in SO(d)$ and

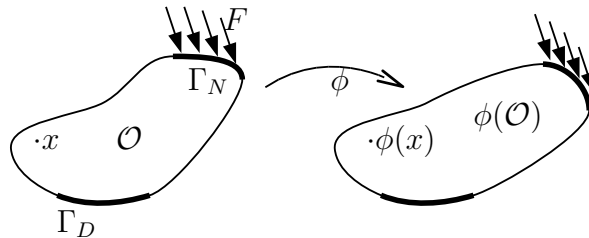


Figure 3.1: A surface load F induces a deformation ϕ of the body \mathcal{O} .

a shift $b \in \mathbb{R}^d$ does not change the energy so that

$$W[\phi] = W(\mathcal{D}\phi) = W(Q^T \mathcal{D}\phi Q) \quad \forall Q \in SO(d).$$

We will furthermore assume an isotropic material so that a rotation of the material before applying a deformation yields the same energy as before,

$$W[\phi] = W(\mathcal{D}\phi) = W(\mathcal{D}\phi Q) \quad \forall Q \in SO(d).$$

The above two conditions lead to the fact that the energy density W only depends on the singular values, the so-called principal stretches, $\lambda_1, \lambda_2, \lambda_3$ of $\mathcal{D}\phi$. Instead of the principal stretches, we can equivalently describe the local deformation using the so-called invariants of the deformation gradient,

$$\begin{aligned} I_1 &= \|\mathcal{D}\phi\|_F = \sqrt{\lambda_1^2 + \lambda_2^2 + \lambda_3^2}, \\ I_2 &= \|\text{cof}\mathcal{D}\phi\|_F = \sqrt{\lambda_1^2 \lambda_2^2 + \lambda_1^2 \lambda_3^2 + \lambda_2^2 \lambda_3^2}, \\ I_3 &= \det \mathcal{D}\phi = \lambda_1 \lambda_2 \lambda_3, \end{aligned}$$

where $\|A\|_F = \sqrt{\text{tr}(A^T A)}$ for $A \in \mathbb{R}^{d \times d}$ and the cofactor matrix is given by $\text{cof}A = \det A A^{-T}$ for $A \in GL(d)$, so that overall, for some appropriate \hat{W} we obtain

$$W[\phi] = W(\mathcal{D}\phi) = \hat{W}(I_1, I_2, I_3).$$

$I_1, I_2,$ and I_3 can be interpreted as the locally averaged change of an infinitesimal length, area, and volume during the deformation, respectively.

The elastic energy stored inside the material, $\mathcal{W}[\mathcal{O}, \phi]$, is thus given by the accumulated energy density (where we assume $W(\mathcal{D}\phi)$ to be measurable in x),

$$\mathcal{W}[\mathcal{O}, \phi] = \int_{\mathcal{O}} W(\mathcal{D}\phi) \, dx.$$

However, there is also a proportion of mechanical energy due to the work of the surface load. By definition, a surface load is the conjugate variable to the displacement $\phi(x) - x$, that is, the surface load $F \, da(x)$ on an infinitesimal area element $da(x) \subset \Gamma_N$ at a point $x \in \Gamma_N$ is the negative rate of change of the mechanical energy for a small change of the displacement $(\phi(x) - x)$ (such a surface load manifests itself as a force density per surface area). Therefore, the external energy term is given by $-\int_{\Gamma_N} F \cdot (\phi - \text{id}) \, da$ so that the overall free energy becomes

$$\mathcal{E}[\mathcal{O}, \phi] = \int_{\mathcal{O}} W(\mathcal{D}\phi) \, dx - \int_{\Gamma_N} F \cdot (\phi - \text{id}) \, da.$$

The thermodynamic equilibrium deformation is given by the minimiser of the variational problem

$$\phi = \arg \min_{\tilde{\phi} \in V} \mathcal{E}[\mathcal{O}, \tilde{\phi}]$$

within a suitable space V of deformations that satisfy $\phi(x) = x$ on Γ_D (the appropriate space V will later turn out to be a Sobolev space $W^{1,p}(\mathcal{O})$ for some $p > 1$). Provided a minimiser ϕ exists, the Euler-Lagrange equation $0 = \langle \partial_\phi \mathcal{E}[\mathcal{O}, \phi], \theta \rangle$ (where $\langle \delta_z \mathcal{G}, \zeta \rangle$ shall denote the Gâteaux derivative of an energy \mathcal{G} with respect to z in some test direction ζ) of the above minimisation yields a (elliptic) partial differential equation in weak form which is solved by ϕ ,

$$0 = \int_{\mathcal{O}} W_{,A}(\mathcal{D}\phi) : \mathcal{D}\theta \, dx - \int_{\Gamma_N} F \cdot \theta \, da \quad \forall (\theta + \text{id}) \in V,$$

where $W_{,A}$ shall denote the derivative of W with respect to its matrix argument and $A : B = \text{tr}(A^T B)$ for $A, B \in \mathbb{R}^{d \times d}$. In its strong form, via integration by parts we obtain

$$\begin{aligned} \text{div } W_{,A}(\mathcal{D}\phi) &= 0 && \text{in } \mathcal{O}, \\ \phi &= \text{id} && \text{on } \Gamma_D, \\ W_{,A}(\mathcal{D}\phi)\nu^{\text{ref}} &= F && \text{on } \Gamma_N, \\ W_{,A}(\mathcal{D}\phi)\nu^{\text{ref}} &= 0 && \text{on } \partial\mathcal{O} \setminus (\Gamma_D \cup \Gamma_N), \end{aligned}$$

where ν^{ref} denotes the unit outward normal on $\partial\mathcal{O}$.

Apparently, $W_{,A}(\mathcal{D}\phi)\nu^{\text{ref}}$ has the interpretation of a surface load, that is, a force per area on the surface perpendicular to ν^{ref} . By cutting \mathcal{O} along a plane with normal ν and then repeating the integration by parts on both sides, we observe that $W_{,A}(\mathcal{D}\phi)\nu$ is the force density acting on the cutting plane. Hence, $\sigma^{\text{ref}} := W_{,A}(\mathcal{D}\phi)$ has the interpretation of a stress tensor, and the actual stress in a plane perpendicular to some unit vector ν is recovered as $\sigma^{\text{ref}}\nu$. With this interpretation, the first equation of the strong formulation represents the conservation of linear momentum, $\text{div } \sigma^{\text{ref}} = 0$, while the corresponding constitutive law, which couples the stress and the deformation, is given by $\sigma^{\text{ref}} = W_{,A}(\mathcal{D}\phi)$.

The above equations are given in a Lagrangian description, that is, stresses are expressed as forces per area in the reference configuration. In more detail, given a force $\hat{F}[dA^{\text{ref}}]$ acting on an area element dA^{ref} with normal ν^{ref} (Figure 3.2, left), the stress is given by

$$\sigma^{\text{ref}}\nu^{\text{ref}} = \lim_{|dA^{\text{ref}}| \rightarrow 0} \frac{\hat{F}[dA^{\text{ref}}]}{|dA^{\text{ref}}|}.$$

This stress is called the first Piola–Kirchhoff stress. The situation can also be described in the deformed configuration, where the force $\hat{F}[dA] = \hat{F}[dA^{\text{ref}}]$ is actually acting, which is then called the Eulerian description. Here, the deformed area element and its normal are denoted dA and ν , and we define the so-called Cauchy stress

$$\sigma\nu = \lim_{|dA| \rightarrow 0} \frac{\hat{F}[dA]}{|dA|}.$$

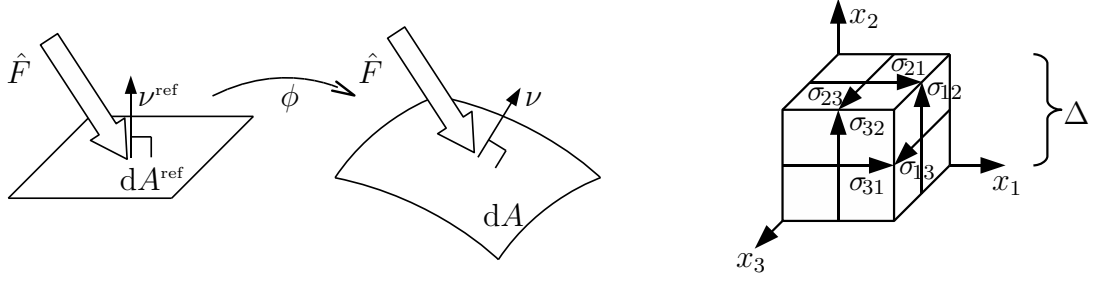


Figure 3.2: Left: Force on an area element in the reference and the deformed configuration. Right: The angular momentum of a material cube with infinitesimal side length Δ around its centre is $\Delta(\sigma_{23} - \sigma_{32}, \sigma_{31} - \sigma_{13}, \sigma_{12} - \sigma_{21})^T$ (normal stresses are not displayed for the sake of clarity).

Via the purely geometric relation $\nu dA = \text{cof} \mathcal{D}\phi \nu^{\text{ref}} dA^{\text{ref}}$ and $\sigma \nu dA = \hat{F}[dA] = \hat{F}[dA^{\text{ref}}] = \sigma^{\text{ref}} \nu^{\text{ref}} dA^{\text{ref}}$ we obtain the relation

$$\sigma \circ \phi \text{cof} \mathcal{D}\phi = \sigma^{\text{ref}},$$

where of course both sides are evaluated at points $x \in \mathcal{O}$. It is now easy to derive from the Lagrangian partial differential equations the strong partial differential equations in Eulerian description,

$$\begin{aligned} \text{div } \sigma &= 0 && \text{in } \phi(\mathcal{O}), \\ \phi &= \text{id} && \text{on } \Gamma_D, \\ \sigma \nu &= F && \text{on } \phi(\Gamma_N), \\ \sigma \nu &= 0 && \text{on } \phi(\partial\mathcal{O} \setminus (\Gamma_D \cup \Gamma_N)). \end{aligned}$$

Furthermore, the absence of local angular momentum in equilibrium implies $0 = \sigma_{ij} - \sigma_{ji}$, $i, j = 1, \dots, d$, (Figure 3.2, right) and hence the symmetry of σ , while σ^{ref} is in general not symmetric.

As already mentioned earlier, hyperelastic energy densities are of the form $W(\mathcal{D}\phi) = \hat{W}(I_1, I_2, I_3)$. Typically, these densities have to fulfil certain conditions. First, we require the identity, $\mathcal{D}\phi = I$, (which corresponds to no deformation) to be the global minimiser. Second, the energy density shall converge to infinity as I_3 , the determinant of the deformation gradient (which describes the volume change), approaches zero or infinity. Negative values of I_3 correspond to local interpenetration of matter and are not allowed at all. Thus, $W(\mathcal{D}\phi) = \hat{W}(I_1, I_2, I_3)$ is strongly nonlinear and can in addition not be convex in the deformation gradient $\mathcal{D}\phi$, since the set of matrices with positive determinant is not even a convex set. This makes the problem of existence of minimisers a slightly subtle one, but it can be treated using the direct method of the calculus of variations: We require coercivity of $\mathcal{E}[\mathcal{O}, \cdot]$ in some convenient topology so that from a minimising sequence, we can extract a convergent subsequence $\phi_i \rightarrow_{i \rightarrow \infty} \phi$. As a second

step, we require $\mathcal{E}[\mathcal{O}, \cdot]$ to be sequentially lower semi-continuous along this sequence so that $\liminf_{i \rightarrow \infty} \mathcal{E}[\mathcal{O}, \phi_i] \geq \mathcal{E}[\mathcal{O}, \phi]$, and hence, ϕ must be a minimiser.

The appropriate topology is the weak topology of a Sobolev space, since via the energy density W we can obtain control over the deformation gradient and then exploit the reflexivity of the Sobolev space to find a weakly convergent subsequence. Hence, we will assume $W(\mathcal{D}\phi) \geq C_1 \|\mathcal{D}\phi\|_F^p - C_2$ for some $p > 1$, $C_1, C_2 > 0$, as well as $F \in L^{p'}(\Gamma_N)$ for $1 = \frac{1}{p} + \frac{1}{p'}$, and we obtain

$$\begin{aligned} \mathcal{E}[\mathcal{O}, \phi] &\geq C_1 \|\mathcal{D}\phi\|_{L^p(\mathcal{O})}^p - C_2 |\mathcal{O}| - \|F\|_{L^{p'}(\Gamma_N)} \|\phi\|_{L^p(\Gamma_N)} + \int_{\Gamma_N} F \cdot x \, da \\ &\geq \hat{C} (\|\phi\|_{W^{1,p}(\mathcal{O})} - \tilde{C})^p - \bar{C} \|\phi\|_{W^{1,p}(\mathcal{O})} - \bar{C} \end{aligned}$$

for some $\hat{C}, \tilde{C}, \bar{C} > 0$, where in the first step we have used Hölder's inequality and in the second step Poincaré's inequality (recall $\phi = \text{id}$ on Γ_D) as well as the boundedness of the trace operator $W^{1,p}(\mathcal{O}) \rightarrow L^p(\Gamma_N)$. Hence, from boundedness of $\mathcal{E}[\mathcal{O}, \phi_i]$ for some sequence ϕ_i , $i \in \mathbb{N}$, we may deduce boundedness of $\|\phi_i\|_{W^{1,p}(\mathcal{O})}$ and thus (by the reflexivity of $W^{1,p}(\mathcal{O})$ and the consistency of the boundary condition $\phi|_{\Gamma_D} = \text{id}$ with weak convergence) the desired weak coercivity of $\mathcal{E}[\mathcal{O}, \cdot]$.

Obviously, for the weak lower semi-continuity of $\mathcal{E}[\mathcal{O}, \cdot]$ we have to require lower semi-continuity of W . Furthermore, it is well-known that in higher dimensions, the weak lower semi-continuity of $\mathcal{W}[\mathcal{O}, \cdot]$ translates into quasiconvexity of W [45], which is, however, difficult to examine. Fortunately, a slightly stronger notion can be applied here. We require W to be polyconvex, that is, there is a convex function $\bar{W} : \mathbb{R}^{d \times d} \times \mathbb{R}^{d \times d} \times \mathbb{R} \rightarrow \mathbb{R}$ with

$$W(A) = \bar{W}(A, \text{cof} A, \det A) \quad \forall A \in \mathbb{R}^{d \times d}.$$

In that case, by a compensated compactness result due to Ball [11], $\mathcal{W}[\mathcal{O}, \cdot]$ is weakly lower semi-continuous on $W^{1,p}(\mathcal{O})$ for $p \geq d$ [18], and we thus have the following result.

Theorem 1. *Let $W : \mathbb{R}^{d \times d} \rightarrow \mathbb{R}$ be polyconvex with $W(A) \geq C_1 \|A\|_F^p - C_2$, $p \geq d$, then the variational problem $\min_{\phi} \mathcal{E}[\mathcal{O}, \phi]$ admits a minimiser in $\{\phi \in W^{1,p}(\mathcal{O}) : \phi|_{\Gamma_D} = \text{id}\}$.*

By imposing growth conditions of the form $W(A) \geq C_1 (\|A\|_F^p + \|\text{cof} A\|_F^q + |\det A|^r) - C_2$ one may even obtain existence results for smaller p under appropriate conditions on q and r [93]. Typical energy densities of the above type are given by

$$W(\mathcal{D}\phi) = \hat{W}(I_1, I_2, I_3) = a_1 \|\mathcal{D}\phi\|_F^p + a_2 \|\text{cof} \mathcal{D}\phi\|_F^q + \Gamma(\det \mathcal{D}\phi)$$

for $a_1, a_2 > 0$, $p, q > 1$, and a convex function $\Gamma : \mathbb{R} \rightarrow \mathbb{R}$ with $\lim_{d \rightarrow \infty} \Gamma(d) = \lim_{d \rightarrow 0} \Gamma(d) = \infty$. For example, $p = 2$ and $q = 0$ yields a neo-Hookean material law, while $p = q = 2$ results in a Mooney–Rivlin material law [38].

Under slightly stronger growth conditions on W , we can obtain additional properties of the minimising deformation. Ball has shown for $p > d$ that if $\phi \in W^{1,p}(\mathcal{O})$ coincides on $\partial\mathcal{O}$ with a homeomorphism of \mathcal{O} and if, for some $\tilde{q} > d$, ϕ satisfies

$$\det \mathcal{D}\phi > 0 \text{ a. e. in } \mathcal{O},$$

$$\int_{\mathcal{O}} \|(\mathcal{D}\phi)^{-1}\|_F^{\tilde{q}} \det \mathcal{D}\phi \, dx < \infty,$$

then ϕ is actually a homeomorphism, and the transformation rule

$$\int_{\mathcal{O}} f \circ \phi \det \mathcal{D}\phi \, dx = \int_{\phi(\mathcal{O})} f \, dx$$

holds for any measurable $f : \mathbb{R}^d \rightarrow \mathbb{R}$ if one of both integrals exists [12, 115]. Hence, if we assume $W(A) = \infty$ for $\det A \leq 0$ and

$$W(A) \geq C_1(\|A\|_F^p + \|\operatorname{cof} A\|_F^q + |\det A|^r + |\det A|^{-s}) - C_2$$

for $p, q > d$, $s > \frac{(d-1)q}{q-d}$, $r > 1$, then for the energy-minimising deformation ϕ we have

$$\varepsilon^{-s} |\{x \in \mathcal{O} : \det \mathcal{D}\phi \leq \varepsilon\}| \leq \int_{\mathcal{O}} (\det \mathcal{D}\phi)^{-s} \, dx \leq \frac{\mathcal{W}[\mathcal{O}, \phi] + C_2 |\mathcal{O}|}{C_1} < \infty,$$

for any $\varepsilon > 0$ and hence $\det \mathcal{D}\phi > 0$ almost everywhere in \mathcal{O} . Furthermore, $\tilde{q} := q \frac{1+s}{s+q} > d$, and by Hölder's inequality,

$$\begin{aligned} \int_{\mathcal{O}} \|(\mathcal{D}\phi)^{-1}\|_F^{\tilde{q}} \det \mathcal{D}\phi \, dx &= \int_{\mathcal{O}} \|\operatorname{cof} \mathcal{D}\phi\|_F^{\tilde{q}} (\det \mathcal{D}\phi)^{1-\tilde{q}} \, dx \\ &\leq \left(\int_{\mathcal{O}} \|\operatorname{cof} \mathcal{D}\phi\|_F^q \, dx \right)^{\frac{\tilde{q}}{q}} \left(\int_{\mathcal{O}} (\det \mathcal{D}\phi)^{q \frac{1-\tilde{q}}{q-\tilde{q}}} \, dx \right)^{\frac{q-\tilde{q}}{q}} \\ &\leq \frac{1}{C_1} (\mathcal{W}[\phi] + C_2 |\mathcal{O}|)^{\frac{\tilde{q}}{q}} (\mathcal{W}[\phi] + C_2 |\mathcal{O}|)^{\frac{q-\tilde{q}}{q}} < \infty \end{aligned}$$

so that the above conditions are satisfied and thus the energy minimising deformation ϕ is a homeomorphism (if $\partial \mathcal{O}$ is suitably deformed) so that interpenetration of matter can be excluded.

For later reference, we give here the formulae for the derivative of W ,

$$\begin{aligned} W_{,A}(A) : B &= \partial_{I_1} \hat{W}(\|A\|_F, \|\operatorname{cof} A\|_F, \det A) \frac{1}{\|A\|_F} A : B + \\ &\quad \partial_{I_2} \hat{W}(\|A\|_F, \|\operatorname{cof} A\|_F, \det A) \frac{1}{\|\operatorname{cof} A\|_F} \operatorname{cof} A : \partial_A \operatorname{cof}(A)(B) + \\ &\quad \partial_{I_3} \hat{W}(\|A\|_F, \|\operatorname{cof} A\|_F, \det A) \partial_A \det(A)(B), \end{aligned}$$

where

$$\begin{aligned} \partial_A \det(A)(B) &= \det(A) \operatorname{tr}(A^{-1} B), \\ \partial_A \operatorname{cof}(A)(B) &= \det(A) \operatorname{tr}(A^{-1} B) A^{-T} - \det(A) A^{-T} B^T A^{-T}. \end{aligned}$$

Sometimes it is useful to linearise the relationship $\sigma^{\text{ref}} := W_{,A}(\mathcal{D}\phi)$ for small displacements $v = \phi - \text{id}$ in order to analyse the impact of infinitesimal volume and length variations. This will yield the classical constitutive law of linearised elasticity,

$$\sigma^{\text{ref}} \doteq \sigma \doteq \lambda \text{tr}\epsilon[v]I + 2\mu\epsilon[v], \quad \epsilon[v] := \frac{1}{2}(\mathcal{D}v + \mathcal{D}v^T),$$

where the Lamé constants λ and μ will depend on the particular form of W at the identity and describe the coefficient of volume and length change penalisation, respectively. We can also design W to fit to given λ and μ : As an example, a straightforward calculation reveals that linearisation of the stress for

$$W(\mathcal{D}\phi) = \frac{\mu}{2}\|\mathcal{D}\phi\|_F^2 + \frac{\lambda}{4}(\det\mathcal{D}\phi)^2 - \left(\mu + \frac{\lambda}{2}\right)\log\det\mathcal{D}\phi - \frac{d\mu}{2} - \frac{\lambda}{4}$$

yields exactly the desired relation.

3.2 Phase field description of free interface and discontinuity problems

A number of variational problems in physics and computer vision is associated with the description and detection of (possibly sharp) interfaces or edges. Two prominent problem classes are given by phase transition or segmentation problems and free discontinuity problems, respectively.

Phase transition problems appear, for instance, during solidification and Ostwald ripening in metal alloys. If a molten alloy solidifies, there are usually two or more energetically preferred phases with different compositions into which the metal decomposes. At the same time, the total interface length between the phases is trying to be minimised. For simplicity, let us assume there are just two phases, described by a piecewise constant function $w : \Omega \rightarrow \{-1, 1\}$ in an open region $\Omega \subset \mathbb{R}^d$ which the metal occupies. The corresponding generic variational problem then is to find the phase partition of Ω with minimal Hausdorff measure of the discontinuity set for a prescribed volume of the phases, that is,

$$\min_w \mathcal{H}^{d-1}(\Omega \cap \partial\{x \in \Omega : w(x) = 1\}) \quad \text{on} \quad \left\{ w : \Omega \rightarrow \{-1, 1\} : \int_{\Omega} w \, dx = C \right\}.$$

Analogously, in a classical segmentation problem we would like to partition for example an image $y^0 : \Omega \rightarrow \mathbb{R}$ into two (or more) regions, also described by a piecewise constant function $w : \Omega \rightarrow \{-1, 1\}$, where the length of its discontinuity set is used as a regulariser. The generic segmentation problem for some spatially dependent partition energy density f is of the form

$$\min_w \mathcal{H}^{d-1}(\Omega \cap \partial\{x \in \Omega : w(x) = 1\}) + \int_{\Omega} f(w) \, dx \quad \text{on} \quad \{w : \Omega \rightarrow \{-1, 1\}\}.$$

$f(w)$ may at each $x \in \Omega$ for instance measure the deviation of the given image y^0 from one of two possible grey values c_1, c_2 that is chosen by the segmentation function w ,

$$f(w) = (w + 1)|y^0 - c_1|^2 + (w - 1)|y^0 - c_2|^2.$$

Free discontinuity problems on the other hand aim at finding a pair (y, K) , where $K \subset \Omega \subset \mathbb{R}^d$ is a (sufficiently smooth) union of hypersurfaces and $y : \Omega \setminus K \rightarrow \mathbb{R}^n$ a (sufficiently smooth) function. The generic variational problem reads

$$\min_{K \subset \Omega, y: \Omega \setminus K \rightarrow \mathbb{R}^n} \mathcal{H}^{d-1}(K) + \int_{\Omega \setminus K} |\nabla y|^2 dx + \int_{\Omega \setminus K} f(y) dx,$$

one particular example being the minimisation of the Mumford–Shah functional,

$$\mathcal{E}_{\text{MS}}[y, K] = \int_{\Omega \setminus K} |\nabla y|^2 + \alpha|y - y^0|^2 dx + \nu \mathcal{H}^{d-1}(K),$$

which tries to extract a piecewise smooth cartoon $y : \Omega \rightarrow \mathbb{R}$ from a given image $y^0 : \Omega \rightarrow \mathbb{R}$. Indeed, the L^2 -difference between y and y^0 ensures that y is a good approximation to y^0 , while the gradient term acts as an edge-preserving smoother. A different application comes from fracture mechanics, where $y : \Omega \rightarrow \mathbb{R}^d$ is a displacement of an elastic body and K , the discontinuity set, the positions where the material breaks.

Discontinuous functions or explicit edge sets are difficult to handle numerically, and thus, methods have been developed in which the interface or discontinuity sets are approximated by a smooth so-called phase field function $u \in W^{1,2}(\Omega)$. In the case of phase transition problems, u approximates the piecewise constant function $w : \Omega \rightarrow \{-1, 1\}$, while in the case of free discontinuity problems it is an approximation to $1 - \chi_K$ for the characteristic function χ_K of K . The width of the smooth edge representation in u scales with a scale parameter ε .

The particular form of the phase fields naturally results from minimising the above-mentioned segmentation and free discontinuity energies, where the $(d - 1)$ -dimensional Hausdorff measure of the edge set is replaced by an approximating functional on $W^{1,2}(\Omega)$. In the segmentation problem, this functional is given by the Modica–Mortola energy [89]

$$\mathcal{L}_{\text{MM}}^\varepsilon[u] = \frac{1}{2} \int_{\Omega} \varepsilon |\nabla u|^2 + \frac{1}{\varepsilon} \Psi(u) dx$$

for a potential $\Psi : \mathbb{R} \rightarrow \mathbb{R}$ with $\Psi(-1) = \Psi(1) = 0$ being the global minima. The edge length $\mathcal{H}^{d-1}(K)$ in the free discontinuity model can be approximated by the Ambrosio–Tortorelli approximation [7],

$$\mathcal{L}_{\text{AT}}^\varepsilon[u] = \frac{1}{2} \int_{\Omega} \varepsilon |\nabla u|^2 + \frac{1}{\varepsilon} (1 - u)^2 dx,$$

while $\int_{\Omega \setminus K} |\nabla y|^2 dx$ is approximated by $\int_{\Omega} u^2 |\nabla y|^2 dx$. The resulting one-dimensional phase field profiles perpendicular to the edge set are sketched in Figure 3.3. That the functionals for small ε indeed approximate the length of the edge set and that the optimal phase field profiles look as in Figure 3.3 can be shown via the concept of Γ -convergence.

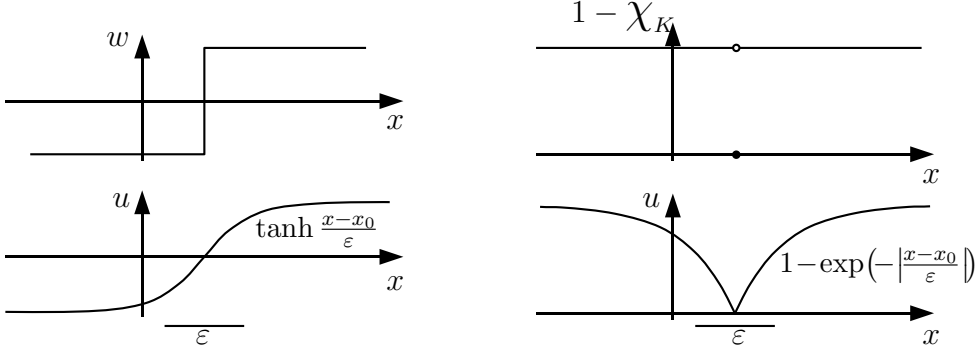


Figure 3.3: Profile of an optimal Modica–Mortola (bottom left, for $\Psi(u) = (1-u^2)^2$) and Ambrosio–Tortorelli phase field (bottom right), approximating a piecewise constant function w and $1 - \chi_K = 1 - \chi_{\{x_0\}}$, respectively.

Definition 2 (Γ -convergence). Let X be a topological space and $f_j : X \rightarrow \mathbb{R} \cup \{\infty\}$, $j \in \mathbb{N}$, a sequence of functionals. We say that f_j Γ -converges to $f : X \rightarrow \mathbb{R} \cup \{\infty\}$, if for every $x \in X$

$$\begin{aligned} \forall x_j \longrightarrow x : f(x) &\leq \liminf_{j \rightarrow \infty} f_j(x_j), \\ \exists x_j \longrightarrow x : f(x) &\geq \limsup_{j \rightarrow \infty} f_j(x_j). \end{aligned}$$

The following theorem about the convergence of minimisers is a direct consequence of this definition (exploiting the existence of converging subsequences for sequences in compact sets, see also [18]).

Theorem 3. If the f_j are equi-mildly coercive on X (that is, there is a compact set $C \subset X$ with $\inf_X f_j = \inf_C f_j$ for all j) and $\Gamma - \lim_{j \rightarrow \infty} f_j = f$, then there exists a minimiser $x \in X$ of f with $f(x) = \lim_{j \rightarrow \infty} \inf_X f_j$. Furthermore, for any precompact sequence x_j such that $\lim_j f_j(x_j) = \lim_{j \rightarrow \infty} \inf_X f_j$, all limit points of that sequence are minimisers of f .

If instead over \mathbb{N} , a family of functionals is parameterised over a parameter ε , which is supposed to converge to zero, Γ -convergence for $\varepsilon \rightarrow 0$ is defined as Γ -convergence for all sequences $\varepsilon_j \rightarrow 0$, and we can use the same notions and techniques.

Concerning the Modica–Mortola functional, let us assume $\Psi \in C^1(\mathbb{R})$ with $\Psi(-1) = \Psi(1) = 0$ being the global minima, then the following result is well-known.

Theorem 4. Define $\mathcal{L}_{\text{MM}}^\varepsilon[u] = \infty$ for $u \notin W^{1,2}(\Omega)$, then

$$\Gamma - \lim_{\varepsilon \rightarrow 0} \mathcal{L}_{\text{MM}}^\varepsilon = \int_{-1}^1 \sqrt{\Psi(s)} \, ds \, \text{Per}(\cdot),$$

where the Γ -limit is taken with respect to the $L^1(\Omega)$ -topology and

$$\text{Per}(w) = \begin{cases} \mathcal{H}^{d-1}(\Omega \cap \partial\{x \in \Omega : w(x) = 1\}) & \text{if } w : \Omega \rightarrow \{-1, 1\}, \\ \infty & \text{else.} \end{cases}$$

A proof can be found in [18]. It is based on finding an optimal profile for a smooth transition between the values -1 and 1 in the one-dimensional setting and then showing this optimal profile to occur perpendicular to the interface by the so-called slicing method. The optimal profile comes from equating both energy contributions in $\mathcal{L}_{\text{MM}}^\varepsilon[u]$ (which results in the minimum possible integrand), yielding an ordinary differential equation

$$u' = \frac{1}{\varepsilon} \sqrt{\Psi(u)}$$

to be solved with boundary conditions $u(-\infty) = -1$ and $u(\infty) = 1$.

For the Ambrosio–Tortorelli functional, the procedure is basically the same, yielding the following [8, 18].

Theorem 5. *Let $\mathcal{E}^\varepsilon[y, u] = \int_\Omega u^2 |\nabla y|^2 dx + \nu \mathcal{L}_{\text{AT}}^\varepsilon[u]$ for $y, u \in W^{1,2}(\Omega)$ and $\mathcal{E}^\varepsilon[y, u] = \infty$ else, then*

$$\Gamma - \lim_{\varepsilon \rightarrow 0} \mathcal{E}^\varepsilon = \mathcal{E}$$

for

$$\mathcal{E}[y, u] = \begin{cases} \int_{\Omega \setminus K[y]} |\nabla y|^2 dx + \nu \mathcal{H}^{d-1}(K[y]) & \text{if } u = 1 \text{ a. e. on } \Omega \\ & \text{and } y \in W^{1,2}(\Omega \setminus K[y]), \\ \infty & \text{else,} \end{cases}$$

where the Γ -limit is taken with respect to the $(L^1(\Omega))^2$ -topology and $K[y]$ is the discontinuity set of the piecewise- $W^{1,2}(\Omega)$ function y .

Here, too, equating the energy contributions in $\mathcal{L}_{\text{AT}}^\varepsilon[u]$ yields the differential equation

$$u' = \frac{1-u}{\varepsilon}$$

to be solved for the optimal phase field profile perpendicular to the edge set with boundary conditions $u(0) = 0$ and $u(-\infty) = u(\infty) = 1$.

4 Elastic shape averaging

The definition and computation of an average belongs to the central tasks when examining the structure of a shape space. In this chapter we will define the average \mathcal{S} of a given number of shapes \mathcal{S}_i , $i = 1, \dots, n$, by imposing an elastic distance $d(\cdot, \cdot)$ on the shape space and minimising $\sum_{i=1}^n d(\mathcal{S}, \mathcal{S}_i)^2$. Recall that we interpret shapes as boundaries $\partial\mathcal{O}$ of deformable objects \mathcal{O} .

Shape averaging has attracted a lot of interest during the past decade, in particular in neuroanatomy research, where standardised anatomical atlases are produced from the anatomy of different patients. By matching these atlases with a patient's data one hopes to detect pathological abnormalities or to be able to precisely locate functional regions [116, 56]. A similar application concerns object recognition by comparison with the average object shape. Shape averages can also be used in manufacturing, for example, for the design of ready-made clothes or shoes that optimally fit the average human stature.

The latter application in particular motivates the use of an elastic distance between shapes despite it not being a metric (compare also Section 1.1). Indeed, we will choose to measure the distance from shape $\mathcal{S}_1 = \partial\mathcal{O}_1$ to $\mathcal{S}_2 = \partial\mathcal{O}_2$ as the square root of the elastic deformation energy needed to deform \mathcal{O}_1 into \mathcal{O}_2 , where we will assume an isotropic, homogeneous, hyperelastic material law. This distance is neither symmetric (the energy to deform \mathcal{O}_2 into \mathcal{O}_1 may be quite different) nor does it satisfy the triangle inequality: In most cases it will cost much less energy to deform \mathcal{O}_1 into some intermediate configuration $\tilde{\mathcal{O}}$ and to additionally deform the object $\tilde{\mathcal{O}}$ into \mathcal{O}_2 . However, these are properties that make an elastic distance appear rather well-suited for the above-mentioned application: As a very crude approximation, assume a shoe to be rigid so that a foot has to be deformed to fit into it. Then the deformation energy is indeed a measure of how similar both shapes are, and this measure should be unidirectional and not necessarily satisfy the triangle inequality in order to strongly weight outliers.

The idea of elastic shape averaging is closely related to groupwise registration. Indeed, our averaging approach can be interpreted as a simultaneous registration of all input shapes with one single (a priori unknown) shape, using an elastic regularisation of the matching deformations. Also, we will extract the shapes from images, using Ambrosio–Tortorelli segmentation. For related registration and segmentation methods as well as different approaches to shape averaging we refer to the review of the literature in Chapter 2.

In the following, we will first propose a sharp interface model for the definition of an elastic shape average in Section 4.1. We will also extend the approach to joint segmen-

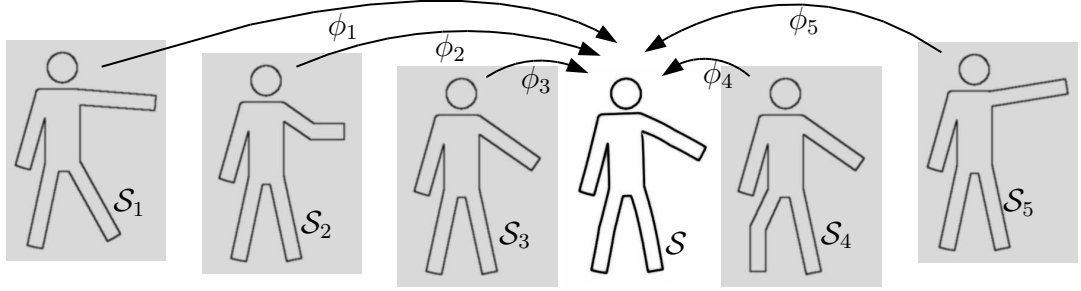


Figure 4.1: Sketch of elastic shape averaging. The input shapes \mathcal{S}_i ($i = 1, \dots, 5$) are mapped onto a shape \mathcal{S} via elastic deformations ϕ_i . The shape \mathcal{S} which minimises the elastic deformation energy is denoted the shape average.

tation and averaging. Subsequently, we will introduce a phase field approximation in Section 4.2. After a brief analysis of the associated existence problem, the numerical implementation is described (Section 4.3) and experimental results are analysed (Section 4.4).

4.1 Shape averages based on nonlinear elastic distances

As emphasised earlier, we employ the notion of shapes as boundaries $\mathcal{S} = \partial\mathcal{O}$ of physically deformable, open bounded objects $\mathcal{O} \subset \mathbb{R}^d$. Now, given a set of shapes, $\mathcal{S}_1, \dots, \mathcal{S}_n$, we seek an average shape \mathcal{S} that reflects the geometric characteristics of the given shapes in a physically intuitive manner. For that purpose it seems generic to interpret the different shapes \mathcal{S}_i and corresponding objects \mathcal{O}_i as deformed configurations of each other. Then, the average shape \mathcal{S} clearly can also be described as a deformed configuration of the input shapes, that is, there are deformations $\phi_i : \mathcal{O}_i \rightarrow \mathbb{R}^d$, $i = 1, \dots, n$, with $\mathcal{S} = \phi_i(\mathcal{S}_i)$ (Figure 4.1). As corresponding average object we obtain $\mathcal{O} = \phi_i(\mathcal{O}_i)$. A natural choice for the definition of the shape average \mathcal{S} then is given by that particular shape \mathcal{S} which minimises the total accumulated deformation energy of all deformations.

This definition of the average as the shape with least elastic deformation energy of the input shapes is related to the arithmetic mean $x = \frac{1}{n} \sum_{i=1}^n x_i$ of given points $x_1, \dots, x_n \in \mathbb{R}^d$. Indeed, x minimises the sum of squared distances, $x = \arg \min_{\tilde{x} \in \mathbb{R}^d} \sum_{i=1}^n |\tilde{x} - x_i|^2$. Given that—by Hooke’s law—the stored elastic energy of an elastic spring extended from x_i to x is proportional to $|x - x_i|^2$, the arithmetic mean x can be interpreted as the minimiser of the total elastic deformation energy in a system where the average x is connected to each x_i by an elastic spring.

4.1.1 Variational definition of the shape average

To be more precise, let $\mathcal{W}[\mathcal{O}, \phi]$ denote the stored elastic deformation energy of a deformation $\phi : \mathcal{O} \rightarrow \mathbb{R}^d$, then the energy to be minimised by the average shape is

$$\mathcal{E}[\mathcal{S}, (\phi_i)_{i=1, \dots, n}] = \begin{cases} \frac{1}{n} \sum_{i=1}^n \mathcal{W}[\mathcal{O}_i, \phi_i] & \text{if } \phi_i(\mathcal{S}_i) = \mathcal{S}, i = 1, \dots, n, \\ \infty & \text{else.} \end{cases}$$

For several reasons, we will consider a nonlinear, hyperelastic deformation energy

$$\mathcal{W}[\mathcal{O}, \phi] = \int_{\mathcal{O}} W(\mathcal{D}\phi) \, dx = \int_{\mathcal{O}} \hat{W}(\|\mathcal{D}\phi\|_F, \|\text{cof}\mathcal{D}\phi\|_F, \det\mathcal{D}\phi) \, dx$$

as introduced in Section 3.1:

- Hyperelastic energies have a physical basis and can be derived from first principles.
- They allow to describe large deformations and thus strong shape variations. In particular, they properly handle isometry-preserving rotations and can therefore capture strong geometric nonlinearities (compare Figure 4.2).
- They allow for material nonlinearities and the distinction between energy changes due to length, area, and volume distortion, which reflect the local distance from an isometry (compare Figure 4.3).
- Under certain conditions they can be used to ensure injectivity of the deformations, resulting in a one-to-one correspondence between the original and the deformed object (compare Section 3.1).

In the following computations, we simply chose $W(\mathcal{D}\phi) = a(\|\mathcal{D}\phi\|_F^2 - d)^2 + b(\det\mathcal{D}\phi^2 + \det\mathcal{D}\phi^{-2})$ for two parameters a and b . It is easily seen that isometries, which satisfy $\|\mathcal{D}\phi\|_F^2 = d$ and $\det\mathcal{D}\phi = 1$, are the global minimisers of this energy density.

Under certain growth conditions on the hyperelastic energy density and if the objects \mathcal{O}_i have a Lipschitz-boundary, we obtain the following ($W^{n,p}$ denotes the Sobolev space of functions with weak derivatives up to order n in L^p).

Theorem 6 (Existence of a shape average). *Let $p > d$ and let the hyperelastic energy density W be polyconvex, have the form $W(A) = \hat{W}(\|A\|_F, \|\text{cof}A\|_F, \det A)$, and satisfy $W(A) \geq C\|A\|_F^p - \tilde{C}$ for some $C, \tilde{C} > 0$ and for all $A \in \mathbb{R}^{d \times d}$. Furthermore, assume there exist homeomorphisms $\psi_{kl} \in W^{1,p}(\mathcal{O}_k)$ between $\overline{\mathcal{O}_k}$ and $\overline{\mathcal{O}_l}$ with $\mathcal{W}[\mathcal{O}_k, \psi_{kl}] < \infty$ for all $1 \leq k, l \leq n$. Then $\mathcal{E}[\mathcal{S}, (\phi_i)_{i=1, \dots, n}]$ admits a minimising shape $\mathcal{S} \subset \mathbb{R}^d$ and corresponding deformations $\phi_i : \mathcal{O}_i \rightarrow \mathbb{R}^d$, $i = 1, \dots, n$.*

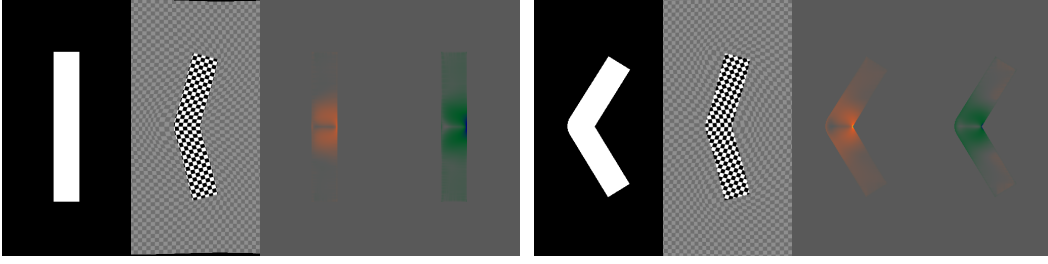



Figure 4.2: A straight and a folded bar as a test case. The input images are depicted along with their deformations ϕ_i (via a deformed checkerboard) and the distribution of the averaged local change of length $\frac{1}{\sqrt{2}}\|\mathcal{D}\phi_i\|_F$ and the local change of area $\det(\mathcal{D}\phi_i)$ with ranges of $[0.97, 1.03]$ colour-coded as . Apparently, isometries are preserved distant from the folding point, and the region of higher deformation energies restricts to the area around the fold. The original bars describe an angle of 180° and 118° , while the average approximately has an angle of 150° . (Resolution 513^2 , $(a, b, \gamma, \eta) = (10^3, 10^3, 1, 10^{-9})$)

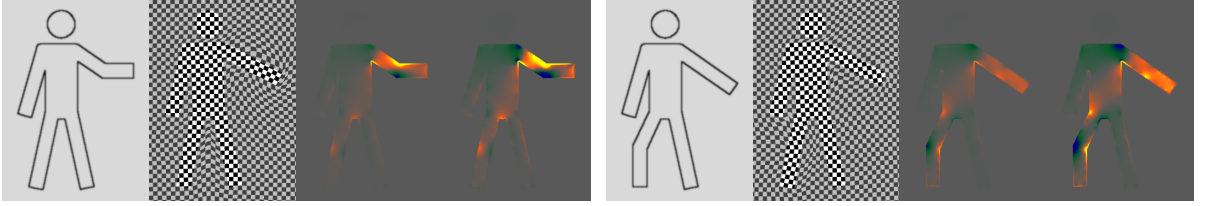



Figure 4.3: For two input shapes from Figure 4.1 the deformation ϕ_i , the averaged local change of length $\frac{1}{\sqrt{2}}\|\mathcal{D}\phi_i\|_F$, and the local change of area $\det\mathcal{D}\phi_i$ are depicted (colours  encode range $[0.95, 1.05]$). (Resolution 513^2 , $(a, b, \gamma, \eta) = (10^3, 10^3, 1, 10^{-9})$)

Proof. The energy \mathcal{E} is not identically infinity due to $\mathcal{E}[\mathcal{S}_l, (\psi_{il})_{i=1, \dots, n}] < \infty$ for $1 \leq l \leq n$. Let a minimising sequence be given by deformations ϕ_i^j and $\mathcal{S}^j = \phi_i^j(\mathcal{S}_i)$, $i = 1, \dots, n$, $j \in \mathbb{N}$. Since the energy is invariant with respect to rigid body motion, we may assume all deformations to have a bounded mean so that Poincaré's inequality may be applied. Due to the growth conditions on W , the ϕ_i^j are uniformly bounded in $W^{1,p}(\mathcal{O}_i)$ and hence weakly converge against a $\phi_i \in W^{1,p}(\mathcal{O}_i)$. By Sobolev embedding, they even converge strongly in $C^{0,\alpha}(\overline{\mathcal{O}}_i)$ for $\alpha < 1 - \frac{d}{p}$. Along this sequence, $\phi_k^j(\mathcal{S}_k) = \mathcal{S}^j = \phi_l^j(\mathcal{S}_l) = \phi_l^j(\psi_{kl}(\mathcal{S}_k))$ for all $1 \leq k, l \leq n$. Furthermore, due to the convergence of the deformations in $C^{0,\alpha}(\overline{\mathcal{O}}_i)$ and the compactness of the shapes \mathcal{S}_i , $\phi_l^j \circ \psi_{kl}(\mathcal{S}_k) = \phi_k^j(\mathcal{S}_k) \rightarrow \phi_k(\mathcal{S}_k)$ for $j \rightarrow \infty$ with respect to the Hausdorff-metric. Also, $\phi_l^j \circ \psi_{kl} \rightarrow \phi_l \circ \psi_{kl}$ in $C^0(\overline{\mathcal{O}}_k)$ and thus $\phi_l^j \circ \psi_{kl}(\mathcal{S}_k) \rightarrow \phi_l \circ \psi_{kl}(\mathcal{S}_k) = \phi_l(\mathcal{S}_l)$ with respect to the Hausdorff-metric. Hence, $\phi_k(\mathcal{S}_k) = \phi_l(\mathcal{S}_l) =: \mathcal{S}$ for all index pairs k, l and thus $\mathcal{E}[\mathcal{S}, (\phi_i)_{i=1, \dots, n}] = \frac{1}{n} \sum_{i=1}^n \mathcal{W}[\mathcal{O}_i, \phi_i]$. However, this energy is weakly lower



Figure 4.4: Deformations in $C^{0,\alpha}$ which (for $\delta \rightarrow 0$) produce arbitrarily large perimeters.

semi-continuous in the deformations due to the polyconvexity of W and $p > d$ (see Section 3.1), hence $\mathcal{E}[\mathcal{S}, (\phi_i)_{i=1,\dots,n}] \leq \liminf_j \mathcal{E}[\mathcal{S}^j, (\phi_i^j)_{i=1,\dots,n}]$ so that a minimiser is given by $(\mathcal{S}, (\phi_i)_{i=1,\dots,n})$. \square

The assumption of input objects with Lipschitz-boundary is essential in the above proof, since otherwise there would be no Sobolev embedding of the deformations into the space of Hölder-continuous functions. In that case, the deformations would not have to be bounded so that we would lack a compactness property for the sequences of deformed shapes $\phi_i^j(\mathcal{S}_i)$ (like the one given by the Blaschke selection theorem) whereas in the above proof, these sequences indeed converge against a \mathcal{S} .

The form of energy $\mathcal{E}[\mathcal{S}, (\phi_i)_{i=1,\dots,n}]$ might unfortunately not be sufficient to ensure enough regularity of the minimising shape \mathcal{S} . In particular, \mathcal{S} might have an arbitrarily large perimeter: The minimising deformations ϕ_i only lie in $W^{1,p}(\mathcal{O}_i)$ and are Hölder-continuous for some Hölder constant $\alpha > 0$ by Sobolev embedding. The regularity theory allows to prove Lipschitz continuity only under certain strong conditions [33]. However, we can deform the unit square $[0, 1]^2$ according to $(x_1, x_2) \mapsto (x_1, x_2(1 + [x_1]_\delta^\alpha))$, where $[x_1]_\delta := |x_1 - \delta - 2\delta \lfloor \frac{x_1}{2\delta} \rfloor|$ (with $\lfloor \cdot \rfloor$ denoting the integer part) shall be a zigzag oscillation of period $2\delta > 0$. This family of deformations, parameterised by $0 < \delta < 1$, is uniformly bounded in $C^{0,\alpha}([0, 1]^2)$, but the perimeter of the deformed unit square tends to infinity as $\delta \rightarrow 0$ (Figure 4.4, left). This particular example does not lie in a Sobolev space $W^{1,p}([0, 1]^2)$, however, one might imagine for instance a similar deformation on a cusped region as in Figure 4.4, right, where the oscillations get faster towards the cusp and which lies in a Sobolev space.

In order to obtain more regularity, we add a regularising prior $\mathcal{L}[\mathcal{S}]$ to the energy which we choose to be the $(d - 1)$ -dimensional Hausdorff measure of \mathcal{S} ,

$$\mathcal{L}[\mathcal{S}] = \mathcal{H}^{d-1}(\mathcal{S}),$$

so that finally, the average shape \mathcal{S} is defined by the variational problem

$$(\mathcal{S}, (\phi_i)_{i=1,\dots,n}) = \arg \min_{\tilde{\mathcal{S}} \subset \mathbb{R}^d, \tilde{\phi}_i: \mathcal{O}_i \rightarrow \mathbb{R}^d} \mathcal{E}[\tilde{\mathcal{S}}, (\tilde{\phi}_i)_{i=1,\dots,n}] + \eta \mathcal{L}[\tilde{\mathcal{S}}]$$

for some small $\eta > 0$. Note that this energy can just as well be used to find the average image morphology \mathcal{S} for given images $y_i: \Omega \rightarrow \mathbb{R}$ with edge sets $\mathcal{S}_i \subset \Omega \subset \mathbb{R}^d$. In this case, the deformations mapping \mathcal{S}_i onto the average edge set \mathcal{S} are simply defined on Ω .

The Euler–Lagrange equations for the minimising deformations ϕ_i result into the system of partial differential equations $\operatorname{div} W_{,\mathcal{A}}(\mathcal{D}\phi_i) = 0$ on \mathcal{O}_i , and a specific condition on

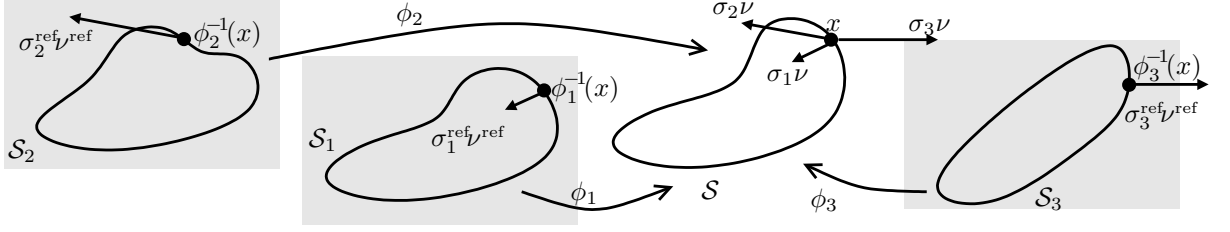


Figure 4.5: Sketch of the pointwise stress balance relation on the averaged shape.

the boundary, expressing the coupling between the deformations: A necessary condition for $\mathcal{S} = \partial\mathcal{O}$ to be a minimiser of $\mathcal{E}[\mathcal{S}, (\phi_i)_{i=1, \dots, n}]$ is that the variation of the energy be zero for any variation of the shape \mathcal{S} and corresponding variations of the deformations ϕ_i . In particular, if we consider a vector field $\delta v : \mathcal{O} \rightarrow \mathbb{R}^d$, then we must have $0 = \frac{d}{d\delta} \mathcal{E}[(\text{id} + \delta v)(\mathcal{S}), ((\text{id} + \delta v) \circ \phi_i)_{i=1, \dots, n}]|_{\delta=0}$. After performing this differentiation, we can integrate by parts to obtain

$$0 = \sum_{i=1}^n \int_{\mathcal{O}_i} W_{,A}(\mathcal{D}\phi_i) : \mathcal{D}(v \circ \phi_i) \, dx = 0 + \sum_{i=1}^n \int_{\mathcal{S}_i} W_{,A}(\mathcal{D}\phi_i) : (v \circ \phi_i) \otimes \nu[\mathcal{S}_i] \, da[\mathcal{S}_i],$$

where $\nu[\mathcal{S}_i]$ is the outer normal in \mathcal{S}_i , and the outer product for two vectors $v_1, v_2 \in \mathbb{R}^d$ is denoted $v_1 \otimes v_2 := v_1 v_2^T$. Recall that $W_{,A}(\mathcal{D}\phi_i)$ represents the first Piola–Kirchhoff stress tensor σ_i^{ref} of the deformation ϕ_i . By considering displacement fields v with local support and letting this support collapse at some point $x \in \mathcal{S}$ we find

$$0 = \sum_{i=1}^n (\sigma_i^{\text{ref}} \nu[\mathcal{S}_i] \, da[\mathcal{S}_i]) (\phi_i^{-1}(x)) \quad \text{and thus} \quad 0 = \sum_{i=1}^n (\sigma_i \nu[\mathcal{S}]) (x),$$

where we have used the relation $(\sigma_i^{\text{ref}} \nu[\mathcal{S}_i] \, da[\mathcal{S}_i]) (\phi_i^{-1}(x)) = (\sigma_i \nu[\mathcal{S}] \, da[\mathcal{S}]) (x)$ between first Piola–Kirchhoff stress and Cauchy stress (see Section 3.1). Hence, the shape average can be interpreted as that stable shape at which the boundary stresses of all deformed input shapes balance each other (Figure 4.5).

4.1.2 A relaxed energy formulation

The hard constraint $\phi_i(\mathcal{S}_i) = \mathcal{S}$ inside the definition of \mathcal{E} is often inadequate in applications. Shapes are usually obtained through some acquisition process which may produce errors or be corrupted by noise. In such cases, different input shapes \mathcal{S}_i might never completely fit to each other so that the constraint has to be relaxed. This can be achieved by adding a mismatch penalty $\mathcal{F}[\mathcal{S}_i, \phi_i, \mathcal{S}]$ to the energy, which is large if $\phi_i(\mathcal{S}_i)$ and \mathcal{S} do not match very well. We choose an edge-based penalty term of the form

$$\mathcal{F}[\mathcal{S}_i, \phi_i, \mathcal{S}] = \mathcal{H}^{d-1}(\mathcal{S}_i \Delta \phi_i^{-1}(\mathcal{S})),$$

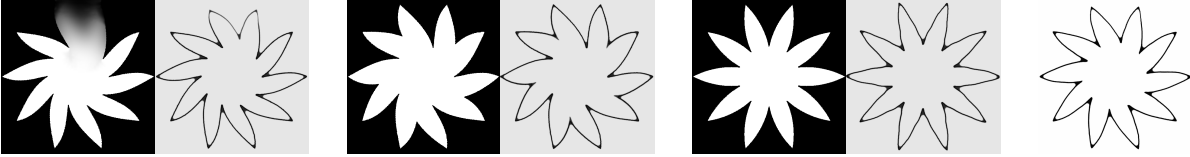


Figure 4.6: Blurred edges can be restored based on a joint approach for image segmentation and averaging. The three input images y_i^0 are depicted along with their segmented edge sets (described as phase field u_i) as computed by the joint segmentation and averaging. The computed average shape is also shown (right). Apparently, the strongly blurred edges in the first input image are reconstructed based on the corresponding edges in the other images. (Resolution 513^2 , $(a, b, \gamma, \eta, \alpha, \beta, \nu) = (10, 10, 1, 10^{-9}, 10^5, 10^{-2}, 10)$)

where $A \Delta B := (A \setminus B) \cup (B \setminus A)$ denotes the symmetric difference between two sets $A, B \subset \mathbb{R}^d$. This type of edge-based penalty is advantageous since it also allows the comparison of image morphologies. The average \mathcal{S} is now defined as the minimiser of

$$\mathcal{E}^\gamma[\mathcal{S}, (\phi_i)_{i=1, \dots, n}] = \frac{1}{n} \sum_{i=1}^n (\mathcal{W}[\mathcal{O}_i, \phi_i] + \gamma \mathcal{F}[\mathcal{S}_i, \phi_i, \mathcal{S}]) + \eta \mathcal{L}[\mathcal{S}]$$

for some $\gamma \gg 0$.

As already mentioned earlier, the averaging problem may be interpreted as a groupwise registration of given shapes $\mathcal{S}_1, \dots, \mathcal{S}_n$ with an a priori unknown shape \mathcal{S} . In this case, \mathcal{F} corresponds to the similarity measure of the registration, while \mathcal{W} acts as a regulariser of the matching deformations.

4.1.3 Joint averaging and segmentation

So far, we have assumed the input shapes $\mathcal{S}_1, \dots, \mathcal{S}_n$ to be given a priori, for example, via previous segmentation of images. However, if the shapes have to be extracted from images, it can be advantageous to simultaneously segment and register or average the shapes. Those image edges which are difficult to extract due to significant noise or low contrast can then be detected by taking into account the corresponding edges in the other input images. This technique has already been applied successfully several times (see references in Section 2.3). While the quality of the registration crucially depends on a robust segmentation result, the segmentation might benefit from the registration since the resulting correspondence between the images might provide complementary information from different images and thus helps to detect weak edges (Figure 4.6).

A joint segmentation and averaging can be achieved via a variational approach by minimising one single functional which contains both the averaging energy as well as a segmentation energy. Since we would like to apply our approach not only to boundaries of open objects \mathcal{O} but also to edge sets within images, the Mumford–Shah energy lends itself

to extract a shape or edge set \mathcal{S} (as well as a piecewise smooth image approximation y) from a given image $y^0 : \Omega \rightarrow \mathbb{R}$ (compare Section 3.2, and see Section 2.3 for references),

$$\mathcal{E}_{\text{MS}}[y, \mathcal{S}, y^0] = \int_{\Omega \setminus \mathcal{S}} |\nabla y|^2 + \alpha |y - y^0|^2 dx + \nu \mathcal{H}^{d-1}(\mathcal{S}).$$

Given input images y_1^0, \dots, y_n^0 , which encode the shapes $\mathcal{S}_1, \dots, \mathcal{S}_n$ to be averaged, we thus define a joint segmentation and averaging energy as

$$\mathcal{E}_{\text{joint}}^\gamma[\mathcal{S}, (y_i, \mathcal{S}_i, \phi_i)_{i=1, \dots, n}] = \frac{1}{n} \sum_{i=1}^n (\beta \mathcal{E}_{\text{MS}}[y_i, \mathcal{S}_i, y_i^0] + \mathcal{W}[\Omega, \phi_i] + \gamma \mathcal{F}[\mathcal{S}_i, \phi_i, \mathcal{S}]) + \eta \mathcal{L}[\mathcal{S}],$$

which is to be minimised simultaneously for the unknowns \mathcal{S} and $y_i, \mathcal{S}_i, \phi_i, i = 1, \dots, n$. In this case, the deformations are naturally defined on the whole of Ω . Figure 4.6 demonstrates that in a joint approach blurry edges in the input images can be segmented if sufficiently strong evidence for these edges from other input images is integrated into the averaged shape.

4.2 Phase field approximation

As described in Section 1.2, there are technical difficulties associated with the use of explicit edge sets \mathcal{S} in a variational setting. We will therefore employ a phase field description of Ambrosio–Tortorelli-type, that is, we will describe a shape \mathcal{S} by a smooth phase field function $u : \Omega \rightarrow \mathbb{R}$ which is zero on \mathcal{S} and close to one everywhere else (see Section 3.2). Such phase fields u can be obtained from images $y^0 : \Omega \rightarrow \mathbb{R}$ with discontinuity set \mathcal{S} by minimising the functional

$$\mathcal{E}_{\text{AT}}^\varepsilon[y, u, y^0] = \int_{\Omega} (u^2 + k_\varepsilon) |\nabla y|^2 dx + \alpha \int_{\Omega} |y - y^0|^2 dx + \frac{\nu}{2} \int_{\Omega} \varepsilon |\nabla u|^2 + \frac{1}{\varepsilon} (u - 1)^2 dx$$

which for small ε approximates the Mumford–Shah functional, where ε can be interpreted as the width of the diffused edge representation in u . In particular, the last integral approximates $\mathcal{H}^{d-1}(\mathcal{S})$ (see Section 3.2). $k_\varepsilon > 0$ is a small parameter that converges to zero as $\varepsilon \rightarrow 0$. It is needed for analytical purposes but will be set to zero in the computations.

The Ambrosio–Tortorelli description of shapes has the advantage that—unlike level set functions or double well phase fields—it can also represent general image morphologies, that is, edge sets of images which do not form the boundary of an open region \mathcal{O} . Also, the fact that the Ambrosio–Tortorelli functional is quadratic in the phase field u will later allow an efficient, direct computation of an average phase field for fixed deformations ϕ_i .

4.2.1 Statement of the averaging energy in terms of phase fields

From now on, we will assume n images $y_1^0, \dots, y_n^0 : \Omega \rightarrow \mathbb{R}$ to be given as input. The Ambrosio–Tortorelli segmentation of these images then yields phase fields u_1, \dots, u_n that represent the shapes to be averaged, and the average shape will also be described by a phase field u . Of course, this requires the averaging energy to be changed accordingly.

The perimeter regularisation $\mathcal{L}[\mathcal{S}]$ will be replaced by the Ambrosio–Tortorelli approximation,

$$\mathcal{L}_{\text{AT}}^\varepsilon[u] = \frac{1}{2} \int_{\Omega} \varepsilon |\nabla u|^2 + \frac{1}{\varepsilon} (u - 1)^2 dx,$$

which for $\varepsilon \rightarrow 0$ Γ -converges against the $(d - 1)$ -dimensional Hausdorff measure of the edge set \mathcal{S} .

The mismatch penalty $\mathcal{F}[\mathcal{S}_i, \phi_i, \mathcal{S}]$ will be replaced by

$$\mathcal{F}^\varepsilon[u_i, \phi_i, u] = \frac{1}{\varepsilon} \int_{\Omega} (u \circ \phi_i)^2 (1 - u_i)^2 + u_i^2 (1 - u \circ \phi_i)^2 dx,$$

where u is supposed to be extended by 1 outside the computational domain Ω . The first term is close to one only near $\mathcal{S}_i \setminus \phi_i^{-1}(\mathcal{S})$, since $u_i \approx 0$ and $u \circ \phi_i \approx 1$ there. Away from that set, it is expected to be close to zero. Analogously, the second term acts as an approximate indicator function of $\phi_i^{-1}(\mathcal{S}) \setminus \mathcal{S}_i$. Let us note that this energy is not expected to be truly proportional to $\mathcal{F}[\mathcal{S}_i, \phi_i, \mathcal{S}]$: First of all, the integrand will (after proper rescaling) only approximate the $(d - 1)$ -dimensional Hausdorff-measure on $(\mathcal{S}_i \setminus \phi_i^{-1}(\mathcal{S})) \cup (\phi_i^{-1}(\mathcal{S}) \setminus \mathcal{S}_i)$ if ϕ_i is neither distending nor compressing perpendicular to the interface. Second, as ε tends to zero, $\mathcal{F}^\varepsilon[u_i, \phi_i, u]$ is bounded from below by some constant times $\mathcal{H}^{d-1}(\mathcal{S}_i)$. Indeed, for a given phase field u_i , the integrand is minimised by $u \circ \phi_i = (1 + (1 - \frac{1}{u_i})^2)^{-1}$, which implies $\mathcal{F}^\varepsilon[u_i, \phi_i, u] \geq \frac{1}{\varepsilon} \int_{\Omega} (\frac{1}{u_i^2} + \frac{1}{(1-u_i)^2})^{-1} dx$. This bound is (almost) proportional to $\mathcal{H}^{d-1}(\mathcal{S}_i)$, and it is (almost) invariant with respect to ε since the width of the edges in u_i scales with ε . Nevertheless, $\mathcal{F}^\varepsilon[u_i, \phi_i, u]$ acts as a proper penalty functional, especially as the latter phenomenon just has the effect of a constant energy offset.

In fact, the structure of the penalty functional \mathcal{F}^ε implies a certain stiffness of the deformations ϕ_i on the diffused interface around the shapes \mathcal{S}_i , since \mathcal{F}^ε tries to match the profiles of the given phase field functions u_i with the pullback $u \circ \phi_i$ of the average phase field u . Indeed, the set of deformations ϕ_1, \dots, ϕ_n tries to minimise stretch or compression perpendicular to the shape contour (Figure 4.7). This does however not hamper the elastic deformation in the limit for $\varepsilon \rightarrow 0$ and $\gamma \rightarrow \infty$, because the other (tangential) components of the deformation can relax freely.

Finally, we have to change the elastic deformation energy $\frac{1}{n} \sum_{i=1}^n \mathcal{W}[\mathcal{O}_i, \phi_i]$: For the mismatch penalty in phase field terms, we apparently need the deformations ϕ_i to be defined everywhere in Ω . For numerical reasons and in order to prevent arbitrarily

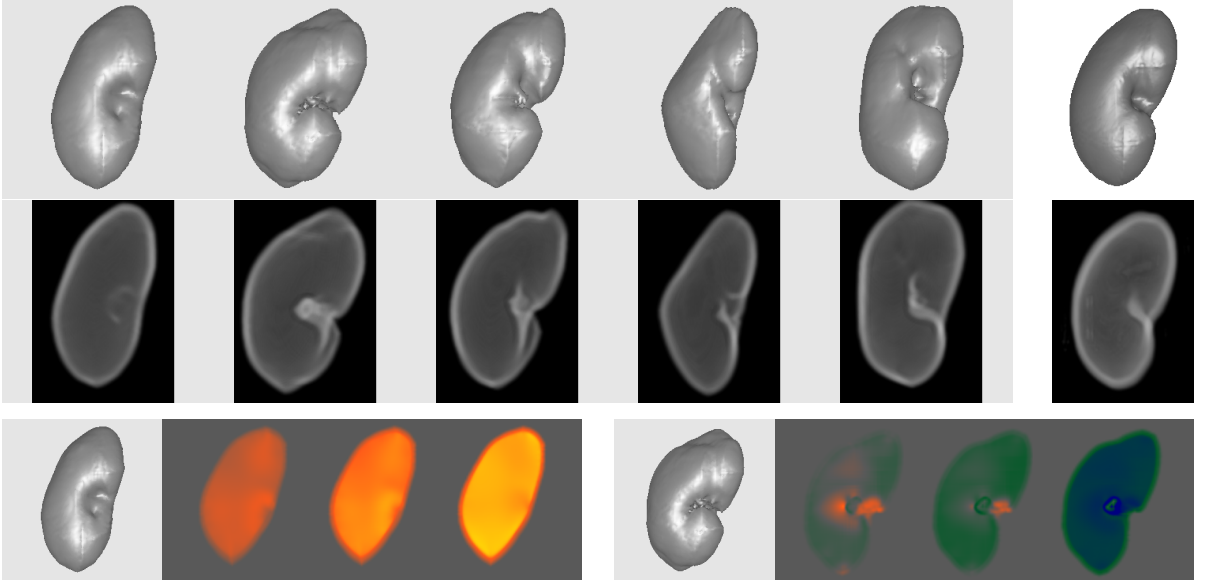


Figure 4.7: Five segmented kidneys and their average (top row). The middle row shows the volume renderings of the corresponding phase fields. The bottom row shows a sagittal cross-section through the distribution of $\frac{1}{\sqrt{3}}\|\mathcal{D}\phi_i\|_F$, $\frac{1}{\sqrt{3}}\|\text{cof}(\mathcal{D}\phi_i)\|_F$, and $\det(\mathcal{D}\phi_i)$ for the first two kidneys (the ranges of $[0.85, 1.15]$ is colour-coded as). While the first kidney is dilated towards the average, the second is compressed. In the thin diffusive interface region, the dilation or compression is reduced. (Resolution 257^3 , $(a, b, \gamma, \eta) = (10, 1, 1, 10^{-7})$)

irregular deformations outside \mathcal{O}_i we will assume the same hyperelastic material law on $\Omega \setminus \mathcal{O}_i$ as on \mathcal{O}_i , but with the stiffness being several orders of magnitude smaller. More precisely, we assume a smooth approximation $\chi_{\mathcal{O}_i}^\varepsilon$ of the characteristic function $\chi_{\mathcal{O}_i}$ to be given based on the prior segmentation, and we define

$$\mathcal{W}^\varepsilon[\mathcal{O}_i, \phi_i] = \int_{\Omega} ((1 - \delta)\chi_{\mathcal{O}_i}^\varepsilon + \delta)\hat{W}(\|\mathcal{D}\phi_i\|_F, \|\text{cof}\mathcal{D}\phi_i\|_F, \det\mathcal{D}\phi_i) dx,$$

where δ was taken to be 10^{-4} in computations. This regularisation, coupled with Dirichlet boundary conditions for the displacement at $\partial\Omega$, also has the effect of ensuring the objects \mathcal{O}_i and shapes \mathcal{S}_i to be deformed homeomorphically so that a material overlap can be ruled out (as will be proven in Section 4.2.3).

Figure 4.8 shows the impact of the choice of the elastic domain on the average shape. Here, we once consider the whole computational domain as homogeneously elastic, and alternatively (and in many cases physically more sound) only the object domain is assumed to be elastic and considerably stiff. The region between both lobes is more severely dilated if the elastic energy is weighted with a small factor outside the shape,

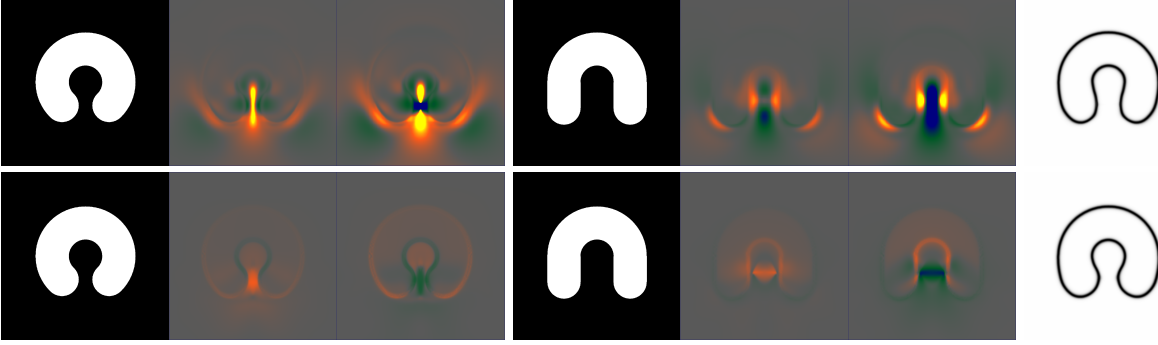



Figure 4.8: Input images together with $\frac{1}{\sqrt{2}}\|\mathcal{D}\phi_i\|_F$ and $\det(\mathcal{D}\phi_i)$ (range of $[0.6, 1.4]$ colour-coded as ) and the average phase field (rightmost). In the top row, only the interior of the two shapes is considerably stiff, whereas the whole computational domain is considered to be homogeneously elastic in the bottom row. Obviously, in the upper case, far stronger strains are visible in the region of the gap, and in the lower case, it is much more expensive to pull the lobes apart in the first shape than to push them together in the second shape. Hence, the resulting average in the second row is characterised by stronger bending of the two lobes than in the first row. (Resolution 1025^2 , $(a, b, \gamma, \eta, \varepsilon) = (0.1, 0.1, 1, 10^{-8}, 6h)$)

which becomes obvious especially in the plots of the deformation invariants.

Overall, the shape averaging functional in terms of phase fields reads

$$\mathcal{E}^{\gamma, \varepsilon}[u, (\phi_i)_{i=1, \dots, n}] = \frac{1}{n} \sum_{i=1}^n (\mathcal{W}^\varepsilon[\mathcal{O}_i, \phi_i] + \gamma \mathcal{F}^\varepsilon[u_i, \phi_i, u]) + \eta \mathcal{L}_{\text{AT}}^\varepsilon[u],$$

which is to be minimised in the average phase field u and deformations ϕ_1, \dots, ϕ_n for given and fixed input phase fields u_1, \dots, u_n . The corresponding energy for joint segmentation and averaging is given by

$$\mathcal{E}_{\text{joint}}^{\gamma, \varepsilon}[u, (y_i, u_i, \phi_i)_{i=1, \dots, n}] = \frac{1}{n} \sum_{i=1}^n (\beta \mathcal{E}_{\text{AT}}^\varepsilon[y_i, u_i, y_i^0] + \mathcal{W}^\varepsilon[\Omega, \phi_i] + \gamma \mathcal{F}^\varepsilon[u_i, \phi_i, u]) + \eta \mathcal{L}_{\text{AT}}^\varepsilon[u],$$

which is to be minimised in u and $y_i, u_i, \phi_i, i = 1, \dots, n$, for given input images y_1^0, \dots, y_n^0 . Note that we are particularly interested in the case where \mathcal{F}^ε acts as a penalty with $\gamma \gg 1$ and $\mathcal{L}_{\text{AT}}^\varepsilon$ ensures a mild regularisation of the averaged shape with $\eta \ll 1$.

4.2.2 Euler–Lagrange equations

A (local) minimum u, ϕ_1, \dots, ϕ_n of $\mathcal{E}^{\gamma, \varepsilon}$ is characterised by the Euler–Lagrange conditions,

$$\langle \delta_u \mathcal{E}^{\gamma, \varepsilon}[u, (\phi_i)_{i=1, \dots, n}], \vartheta \rangle = 0, \quad \langle \delta_{\phi_i} \mathcal{E}^{\gamma, \varepsilon}[u, (\phi_j)_{j=1, \dots, n}], \theta \rangle = 0,$$

where $\langle \delta_z \mathcal{G}, \zeta \rangle$ shall denote the Gâteaux derivative of an energy \mathcal{G} with respect to z in some test direction ζ , and ϑ and θ are scalar- and vector-valued test functions, respectively. These derivatives are given by $\delta_u \mathcal{E}^{\gamma, \varepsilon} = \frac{\gamma}{n} \sum_{i=1}^n \delta_u \mathcal{F}^\varepsilon + \eta \delta_u \mathcal{L}_{\text{AT}}^\varepsilon$ and $\delta_{\phi_i} \mathcal{E}^{\gamma, \varepsilon} = \frac{1}{n} \sum_{i=1}^n (\delta_{\phi_i} \mathcal{W}^\varepsilon + \gamma \delta_{\phi_i} \mathcal{F}^\varepsilon)$, where for sufficiently smooth u and ϕ_i we have

$$\begin{aligned} \langle \delta_u \mathcal{F}^\varepsilon[u_i, \phi_i, u], \vartheta \rangle &= \frac{2}{\varepsilon} \int_{\phi_i(\Omega)} (u(1 - u_i \circ \phi_i^{-1})^2 - (u_i \circ \phi_i^{-1})^2(1 - u)) \vartheta |\det(\mathcal{D}\phi_i^{-1})| dx, \\ \langle \delta_{\phi_i} \mathcal{F}^\varepsilon[u_i, \phi_i, u], \theta \rangle &= \frac{2}{\varepsilon} \int_{\Omega} ((1 - u_i)^2(u \circ \phi_i) - u_i^2(1 - u \circ \phi_i)) (\nabla u \circ \phi_i) \cdot \theta dx, \\ \langle \delta_u \mathcal{L}_{\text{AT}}^\varepsilon[u], \vartheta \rangle &= \int_{\Omega} \varepsilon \nabla u \cdot \nabla \vartheta + \frac{1}{\varepsilon} (u - 1) \vartheta dx, \\ \langle \delta_{\phi_i} \mathcal{W}^\varepsilon[\mathcal{O}_i, \phi_i], \theta \rangle &= \int_{\Omega} \left((1 - \delta) \chi_{\mathcal{O}_i}^\varepsilon + \delta \right) W_{,A}(\mathcal{D}\phi_i) : \mathcal{D}\theta dx. \end{aligned}$$

Upon integration by parts and use of the fundamental lemma of the calculus of variations, the Euler–Lagrange equations translate into a coupled system of partial differential equations and corresponding boundary conditions for u and ϕ_1, \dots, ϕ_n . As natural boundary conditions we obtain homogeneous Neumann boundary conditions for u , $\nabla u \cdot \nu[\partial\Omega] = 0$ on $\partial\Omega$, as well as $W_{,A}(\mathcal{D}\phi_i) \nu[\partial\Omega] = \sigma_i^{\text{ref}} \nu[\partial\Omega] = 0$ on $\partial\Omega$, where $\nu[\partial\Omega]$ shall denote the unit outward normal to $\partial\Omega$. The first condition implies that the phase field interface (and thus the shape \mathcal{S} in the limit $\varepsilon \rightarrow 0$) can meet the boundary $\partial\Omega$ only perpendicular to it. The second condition means a tension-free boundary. The really physically relevant boundary condition for the deformation ϕ_i is diffused in the transition layer around $\partial\mathcal{O}_i$, where we obtain the PDE

$$\sigma_i^{\text{ref}} \nabla \chi_{\mathcal{O}_i}^\varepsilon = - \frac{(1 - \delta) \chi_{\mathcal{O}_i}^\varepsilon + \delta}{1 - \delta} \operatorname{div} \sigma_i^{\text{ref}} + \frac{1}{\varepsilon} \kappa^\delta [u_i, \phi_i, u] \nu$$

with $\kappa^\delta [u_i, \phi_i, u] = \frac{2\gamma}{1-\delta} ((1 - u_i)^2(u \circ \phi_i) - u_i^2(1 - u \circ \phi_i)) |\nabla u \circ \phi_i|$ and $\nu = \frac{\nabla u \circ \phi_i}{|\nabla u \circ \phi_i|}$. For smooth shapes \mathcal{S}_i we expect the first summand on the right-hand side to be uniformly bounded in ε and δ , whereas for $\varepsilon, \delta \rightarrow 0$ the scaled gradient of the smoothed characteristic function $\varepsilon \nabla \chi_{\mathcal{O}_i}^\varepsilon$ converges to the normal $\nu[\mathcal{S}_i]$ in the sense of measures, and ν converges to $\nu[\mathcal{S}]$. Thus, in the limit we recover an effective boundary condition $\sigma_i^{\text{ref}} \nu[\mathcal{S}_i] = \kappa^{\delta \rightarrow 0} [u_i, \phi_i, u] \nu[\mathcal{S}]$ on $\partial\mathcal{O}_i$ for every $i = 1, \dots, n$, which is interlinked with the corresponding boundary conditions for the other deformations via the PDE for u .

Finally, for the joint averaging and segmentation model $\mathcal{E}_{\text{joint}}^{\gamma, \varepsilon}[u, (y_i, u_i, \phi_i)_{i=1, \dots, n}]$ we obtain the additional Euler–Lagrange conditions

$$\langle \delta_{u_i} \mathcal{E}_{\text{joint}}^{\gamma, \varepsilon}, \vartheta \rangle = 0, \quad \langle \delta_{y_i} \mathcal{E}_{\text{joint}}^{\gamma, \varepsilon}, \vartheta \rangle = 0$$

with $\delta_{u_i} \mathcal{E}_{\text{joint}}^{\gamma, \varepsilon} = \frac{1}{n} \sum_{i=1}^n (\beta \delta_{u_i} \mathcal{E}_{\text{AT}}^\varepsilon + \gamma \delta_{u_i} \mathcal{F}^\varepsilon)$ and $\delta_{y_i} \mathcal{E}_{\text{joint}}^{\gamma, \varepsilon} = \frac{1}{n} \sum_{i=1}^n \beta \delta_{y_i} \mathcal{E}_{\text{AT}}^\varepsilon$, where

$$\begin{aligned} \langle \delta_{u_i} \mathcal{E}_{\text{AT}}^\varepsilon [y_i, u_i], \vartheta \rangle &= \int_{\Omega} 2 u_i \vartheta |\nabla y_i|^2 + \nu \left(\varepsilon \nabla u_i \cdot \nabla \vartheta + \frac{1}{\varepsilon} (u_i - 1) \vartheta \right) dx, \\ \langle \delta_{u_i} \mathcal{F}^\varepsilon [u_i, \phi_i, u], \vartheta \rangle &= \frac{2}{\varepsilon} \int_{\Omega} ((u \circ \phi_i)^2 (u_i - 1) + u_i (1 - u \circ \phi_i)^2) \vartheta dx, \\ \langle \delta_{y_i} \mathcal{E}_{\text{AT}}^\varepsilon [y_i, u_i], \vartheta \rangle &= 2 \int_{\Omega} \alpha (y_i - y_i^0) \vartheta + u_i^2 \nabla y_i \cdot \nabla \vartheta dx \end{aligned}$$

for any scalar test function ϑ .

4.2.3 Existence of minimisers

In the following, we will show minimisers u, ϕ_1, \dots, ϕ_n of $\mathcal{E}^{\gamma, \varepsilon}[u, (\phi_i)_{i=1, \dots, n}]$ to exist for $\varepsilon, \delta > 0$ fixed and for given input phase fields u_1, \dots, u_n . As explained before, the phase fields u_i represent shapes \mathcal{S}_i , whose average \mathcal{S} is described by the phase field u . We will also show the existence of minimisers for the joint averaging and segmentation model $\mathcal{E}_{\text{joint}}^{\gamma, \varepsilon}[u, (y_i, u_i, \phi_i)_{i=1, \dots, n}]$, where we are only given input images y_1^0, \dots, y_n^0 and where the corresponding phase fields u_1, \dots, u_n are a result of the minimisation.

Theorem 7 (Existence of a phase field shape average). *Let $\Omega \subset \mathbb{R}^d$ have a Lipschitz boundary, $d \in \{2, 3\}$, $\varepsilon, \delta, \gamma, \eta > 0$, and consider the set of admissible deformations $\mathcal{A} := \{\phi : \Omega \rightarrow \Omega : \phi|_{\partial\Omega} = \text{id}\}$. Furthermore, let the integrand W of $\mathcal{W}^\varepsilon[\mathcal{O}_i, \phi_i]$ be polyconvex and satisfy the growth condition $W(A) \geq C_1(\|A\|_F^p + \|\text{cof} A\|_F^q + |\det A|^r + |\det A|^{-s}) - C_2$ for some $C_1, C_2 > 0$ and all $A \in \mathbb{R}^{d \times d}$ with $p, q > d$, $r > 1$, and $s > \frac{(d-1)q}{q-d}$. If the input phase fields $(u_i)_{i=1, \dots, n}$ lie in $W^{1,2}(\Omega)$ with $0 \leq u_i \leq 1$, then the energy*

$$\mathcal{E}^{\gamma, \varepsilon}[u, (\phi_i)_{i=1, \dots, n}] = \frac{1}{n} \sum_{i=1}^n (\mathcal{W}^\varepsilon[\mathcal{O}_i, \phi_i] + \gamma \mathcal{F}^\varepsilon[u_i, \phi_i, u]) + \eta \mathcal{L}_{\text{AT}}^\varepsilon[u]$$

attains its minimum over phase fields u in $W^{1,2}(\Omega)$ and n -tuples $(\phi_i)_{i=1, \dots, n}$ of deformations in $(\mathcal{A} \cap W^{1,p}(\Omega))^n$. Furthermore, the minimising u and ϕ_i , $i = 1, \dots, n$, satisfy $u \in C^{1, \tilde{\alpha}}(\bar{\Omega})$, $\phi_i \in C^{0, \tilde{\beta}}(\bar{\Omega})$, $u \circ \phi_i \in C^{0, \tilde{\beta}}(\bar{\Omega})$ for all $0 < \tilde{\alpha} < 1 - \frac{d}{s+1}$, $0 < \tilde{\beta} < 1 - \frac{d}{p}$. Finally, the minimising deformations are homeomorphisms.

Proof. Apparently, the total energy is bounded from below by zero. Also, $u \equiv 0$ and $\phi_i \equiv \text{id}$, $i = 1, \dots, n$, show that there are phase fields and deformations for which the energy is finite.

Let $((\phi_i^k)_{i=1, \dots, n}, u^k)_{k \in \mathbb{N}}$ be a minimising sequence. In the following, we will frequently replace minimising sequences by subsequences without explicit subsequence indexing.

As shown in Section 3.1, the growth condition on W implies that ϕ_i^k is a homeomorphism with $\det \mathcal{D}\phi_i^k > 0$ almost everywhere, and that the transformation

$$\int_{\Omega} f \circ \phi_i^k \det \mathcal{D}\phi_i^k \, dx = \int_{\Omega} f \, dx$$

holds if any of both integrals exists. Also, for any bounded subset $\mathcal{B} \subset W^{1,t}(\Omega)$, $\mathcal{B}_{\phi_i^k} := \{f \circ \phi_i^k : f \in \mathcal{B}\}$ is precompact in $L^{\tilde{t}}(\Omega)$ for all $\tilde{t} < \frac{s}{s+1} \left(\frac{1}{t} - \frac{1}{d}\right)^{-1}$ since $f \circ \phi_i^k$ is integrable for a homeomorphism ϕ_i^k , and by the above transformation rule and Hölder's inequality,

$$\begin{aligned} \|f \circ \phi_i^k\|_{L^{\tilde{t}}}^{\tilde{t}} &= \int_{\Omega} |f \circ \phi_i^k|^{\tilde{t}} \, dx = \int_{\Omega} |f|^{\tilde{t}} \frac{1}{\det \mathcal{D}\phi_i^k \circ (\phi_i^k)^{-1}} \, dx \\ &\leq \left(\int_{\Omega} |f|^{\tilde{t} \frac{s+1}{s}} \, dx \right)^{\frac{s}{s+1}} \left(\int_{\Omega} \frac{1}{(\det \mathcal{D}\phi_i^k \circ (\phi_i^k)^{-1})^{s+1}} \, dx \right)^{\frac{1}{s+1}} = \|f\|_{L^{\tilde{t} \frac{s+1}{s}}}^{\tilde{t}} \left(\int_{\Omega} (\det \mathcal{D}\phi_i^k)^{-s} \, dx \right)^{\frac{1}{s+1}}, \end{aligned}$$

where $W^{1,t}(\Omega)$ is compactly embedded in $L^{\tilde{t} \frac{s+1}{s}}$ and the last integral is bounded due to the growth condition on W .

Now we construct a different minimising sequence, still denoted $((\phi_i^k)_{i=1,\dots,n}, u^k)_{k \in \mathbb{N}}$, by letting $u^k = u[(\phi_i^k)_{i=1,\dots,n}]$ be the minimiser of $\mathcal{E}^{\gamma,\varepsilon}[\cdot, (\phi_i^k)_{i=1,\dots,n}]$. However, we first have to verify the existence of such a minimiser u^k : For given $(\phi_i^k)_{i=1,\dots,n}$, let $(u^{j,k})_{j \in \mathbb{N}}$ be a minimising sequence of $\mathcal{E}^{\gamma,\varepsilon}[\cdot, (\phi_i^k)_{i=1,\dots,n}]$. Due to the boundedness of $\mathcal{L}_{\text{AT}}^{\varepsilon}[u^{j,k}]$, $u^{j,k}$ is bounded in $W^{1,2}(\Omega)$ so that, for a subsequence, $u^{j,k} \rightharpoonup u^k$ in $W^{1,2}(\Omega)$ for some $u^k \in W^{1,2}(\Omega)$. Since the weak lower semi-continuity of $\mathcal{L}_{\text{AT}}^{\varepsilon}$ is obvious, for u^k to be the minimiser it only remains to prove weak lower semi-continuity of $\mathcal{F}^{\varepsilon}$ as $u^{j,k} \rightharpoonup u^k$. We have

$$\begin{aligned} \mathcal{F}^{\varepsilon}[u_i, \phi_i^k, u^{j,k}] &= \mathcal{F}^{\varepsilon}[u_i, \phi_i^k, (u^{j,k} - u^k) + u^k] = \mathcal{F}^{\varepsilon}[u_i, \phi_i^k, u^k] \\ &+ \int_{\Omega} (u_i^2 + (1-u_i)^2) ((u^k - u^{j,k}) \circ \phi_i^k)^2 \, dx + 2 \int_{\Omega} (u_i^2 (u^k - 1) \circ \phi_i^k + (1-u_i)^2 u^k \circ \phi_i^k) (u^{j,k} - u^k) \circ \phi_i^k \, dx. \end{aligned}$$

The second term is greater than or equal to zero, while for a subsequence the final term converges to zero, as it follows from the following Hölder estimate,

$$\begin{aligned} &\left| \int_{\Omega} (u_i^2 (u^k - 1) \circ \phi_i^k + (1-u_i)^2 u^k \circ \phi_i^k) (u^{j,k} - u^k) \circ \phi_i^k \, dx \right| \\ &\leq \left(\|u_i\|_{L^6(\Omega)}^2 \|u^k \circ \phi_i^k - 1\|_{L^3(\Omega)} + \|1 - u_i\|_{L^6(\Omega)}^2 \|u^k \circ \phi_i^k\|_{L^3(\Omega)} \right) \|(u^{j,k} - u^k) \circ \phi_i^k\|_{L^3(\Omega)}. \end{aligned}$$

By Sobolev embedding, u_i is uniformly bounded in $L^6(\Omega)$. Furthermore, we can apply our above compactness argument for $t = 2$ and $\tilde{t} = 3$ to u^k and $u^{j,k} - u^k$ so that the

right hand side converges to zero for a subsequence.

The phase fields u^k satisfy certain regularity properties. First of all, we observe

$$\mathcal{E}^{\gamma,\varepsilon}[u^k, (\phi_i^k)_{i=1,\dots,n}] \geq \mathcal{E}^{\gamma,\varepsilon}[\max(0, \min(1, u^k)), (\phi_i^k)_{i=1,\dots,n}].$$

Since u^k is already a minimiser, this implies that the sequence $(u^k)_{k \in \mathbb{N}}$ is uniformly bounded by $0 \leq u^k \leq 1$. Furthermore, u^k satisfies the Euler–Lagrange equation

$$-\varepsilon \eta \Delta u^k = -\frac{\eta}{\varepsilon}(u^k - 1) - \frac{2\gamma}{n\varepsilon} \sum_{i=1}^n (u^k (1 - u_i \circ (\phi_i^k)^{-1})^2 - (u_i \circ (\phi_i^k)^{-1})^2 (1 - u^k)) |\det \mathcal{D}(\phi_i^k)^{-1}|$$

in a weak sense. Due to $0 \leq u^k, u_i \leq 1$ and since $\det \mathcal{D}(\phi_i^k)^{-1}$ is uniformly bounded in $L^{s+1}(\Omega)$ (due to the above-mentioned transformation rule for $f \equiv |\det \mathcal{D}(\phi_i^k)^{-1}|^{s+1} = |(\det \mathcal{D}\phi_i^k \circ (\phi_i^k)^{-1})^{-1}|^{s+1}$), the right-hand side is uniformly bounded in $L^{s+1}(\Omega)$, and applying classical elliptic regularity theory [66] we observe that $(u^k)_{k \in \mathbb{N}}$ is uniformly bounded in $W^{2,s+1}(\Omega)$.

Next, we consider the functional

$$(\phi_i)_{i=1,\dots,n} \mapsto \mathcal{E}^{\gamma,\varepsilon}[u[(\phi_i)_{i=1,\dots,n}], (\phi_i)_{i=1,\dots,n}].$$

Due to the growth condition on W , $((\mathcal{D}\phi_i^k, \text{cof}\mathcal{D}\phi_i^k, \det\mathcal{D}\phi_i^k))_{k \in \mathbb{N}}$ is uniformly bounded in $L^p(\Omega) \times L^q(\Omega) \times L^r(\Omega)$. By Poincaré’s inequality applied to $(\phi_i^k - \text{id})$ we obtain that $(\phi_i^k)_{k \in \mathbb{N}}$ is uniformly bounded in $W^{1,p}(\Omega)$, and we can extract a weakly convergent subsequence. By Ball’s classical compensated compactness result [11] we observe that

$$(\mathcal{D}\phi_i^k, \text{cof}\mathcal{D}\phi_i^k, \det\mathcal{D}\phi_i^k) \rightharpoonup (\mathcal{D}\phi_i, \text{cof}\mathcal{D}\phi_i, \det\mathcal{D}\phi_i)$$

in $L^p(\Omega) \times L^q(\Omega) \times L^r(\Omega)$ for some $\phi_i \in \mathcal{A} \cap W^{1,p}(\Omega)$, $i = 1, \dots, n$, and $\mathcal{W}^\varepsilon[\mathcal{O}_i, \cdot]$ is sequentially weakly lower semi-continuous (see Section 3.1). Furthermore, by Sobolev embedding and upon extracting a subsequence, we observe $u^k \rightarrow u$ for some u in $C^{1,\tilde{\alpha}}(\bar{\Omega})$ with $0 \leq \tilde{\alpha} < 1 - \frac{d}{s+1}$ (note that $s > d - 1$) so that $\mathcal{L}_{\text{AT}}^\varepsilon[u^k]$ converges to $\mathcal{L}_{\text{AT}}^\varepsilon[u]$. Finally, we justify the continuity of \mathcal{F}^ε on a subsequence, that is,

$$\mathcal{F}^\varepsilon[u_i, \phi_i^k, u[(\phi_i^k)_{i=1,\dots,n}]] \rightarrow \mathcal{F}^\varepsilon[u_i, \phi_i, u]$$

as $k \rightarrow \infty$: Due to Sobolev’s embedding theorem, a subsequence of ϕ_i^k converges strongly in $C^{0,\tilde{\beta}}(\bar{\Omega})$ for $\tilde{\beta} < 1 - \frac{d}{p}$. For $x, x+z \in \bar{\Omega}$ we estimate

$$|(u^k \circ \phi_i^k)(x+z) - (u^k \circ \phi_i^k)(x)| \leq \|u^k\|_{C^{0,1}(\bar{\Omega})} \|\phi_i^k\|_{C^{0,\tilde{\beta}}(\bar{\Omega})} |z|^{\tilde{\beta}}$$

so that the concatenation of u^k and ϕ_i^k is uniformly Hölder continuous for a positive Hölder exponent $\tilde{\beta}$. By the Arzelà–Ascoli theorem we establish the uniform convergence of $u^k \circ \phi_i^k$ against $u \circ \phi_i$ for another subsequence and thence the requested continuity of the

penalty functional \mathcal{F}^ε (for example, using Lebesgue's dominated convergence theorem). Altogether, we have obtained the sequential weak lower semi-continuity of

$$(\phi_i)_{i=1,\dots,n} \mapsto \mathcal{E}^{\gamma,\varepsilon}[u[(\phi_i)_{i=1,\dots,n}], (\phi_i)_{i=1,\dots,n}],$$

which implies that u is a minimising phase field and $(\phi_i)_{i=1,\dots,n}$ is a set of minimising elastic deformations. Furthermore, the homeomorphism property is a direct consequence of the boundedness of the elastic energy (see above). \square

For the joint energy model, we obtain an analogous result.

Theorem 8 (Existence of minimisers for the joint model). *Let $\alpha, \beta, \nu > 0$ and let the assumptions of the previous theorem hold. Furthermore, let $y_i^0 \in L^2(\Omega)$ for $i = 1, \dots, n$. Then, the energy*

$$\mathcal{E}_{\text{joint}}^{\gamma,\varepsilon}[u, (y_i, u_i, \phi_i)_{i=1,\dots,n}] = \frac{1}{n} \sum_{i=1}^n (\beta \mathcal{E}_{\text{AT}}^\varepsilon[y_i, u_i, y_i^0] + \mathcal{W}^\varepsilon[\Omega, \phi_i] + \gamma \mathcal{F}^\varepsilon[u_i, \phi_i, u]) + \eta \mathcal{L}_{\text{AT}}^\varepsilon[u]$$

attains its minimum over n -tuples of images $y_i \in W^{1,2}(\Omega)$, phase fields $u_i \in W^{1,2}(\Omega)$, and deformations $\phi_i \in \mathcal{A}$ with $i = 1, \dots, n$, and over phase fields $u \in W^{1,2}(\Omega)$. The minimising u , u_i , and ϕ_i satisfy $u_i \in C^{1,\hat{\alpha}}(\bar{\Omega})$, $u_i \in C^{1,\hat{\alpha}}(\bar{\Omega})$, $\phi_i \in C^{0,\hat{\beta}}(\bar{\Omega})$, $u \circ \phi_i \in C^{0,\hat{\beta}}$ for all $0 < \hat{\alpha} < 1$, $0 < \hat{\alpha} < 1 - \frac{d}{s+1}$, $0 < \hat{\beta} < 1 - \frac{d}{p}$. Furthermore, the minimising deformations are homeomorphisms.

Proof. The required arguments are closely related to those in the proof of the previous theorem. Let $((y_i^k, u_i^k, \phi_i^k)_{i=1,\dots,n}, u^k)_{k \in \mathbb{N}}$ be a minimising sequence, where we assume that for fixed $(y_i^k, \phi_i^k)_{i=1,\dots,n}, u^k$, the n -tuple of phase fields $(u_i^k)_{i=1,\dots,n}$ is a minimiser over all n -tuples of phase fields in $W^{1,2}(\Omega)$. The existence of these phase fields is straightforward, and once more by truncation we observe that $0 \leq u_i^k \leq 1$ (and that, analogously to the argument for u^k in the previous proof, u_i^k converges strongly in $C^{1,\hat{\alpha}}$ for all $0 < \hat{\alpha} < 1$). Hence, u_i^k is now an admissible phase field for the description of the input shapes in the previous theorem. Thus, we can again modify the minimising sequence and suppose that for fixed $(u_i^k)_{i=1,\dots,n}$ the other components $(y_i^k, \phi_i^k)_{i=1,\dots,n}, u^k$ minimise the global energy. To prove this, we follow exactly the above proof and remark that the weak coercivity and weak lower semi-continuity of $\mathcal{E}_{\text{AT}}^\varepsilon[\cdot, u_i^k]$ are obvious. Finally, we repeat the arguments to obtain a (weak) limit of $((y_i^k, u_i^k, \phi_i^k)_{i=1,\dots,n}, u^k)_{k \in \mathbb{N}}$ as well as the sequential lower semi-continuity of the total energy as $k \rightarrow \infty$ (for the continuity of \mathcal{F}^ε we again use Lebesgue's dominated convergence theorem and the pointwise convergence of $u^k \circ \phi_i^k$ and u_i^k), which concludes the proof. \square

4.3 Numerical implementation

The problem is discretised using continuous multilinear finite elements on a regular grid. In detail, we consider $\Omega = [0, 1]^d$ and overlay Ω with a cubic lattice of $2^L + 1$ equispaced

vertices in each space direction (and thus $(2^L + 1)^d$ vertices in total), which corresponds to a grid size of $h = 2^{-L}$. The images y_i , phase fields u, u_i , and deformations ϕ_i will be represented by continuous, piecewise multilinear (bilinear for $d = 2$ and trilinear for $d = 3$) finite element functions, where each pixel or voxel corresponds to a mesh node.

Let I_h denote the index set of all vertices, then each vertex $x_i, i \in I_h$, is associated with a continuous, piecewise multilinear hat function φ_i satisfying $\varphi_i(x_j) = 0$ for $i \neq j$ and $\varphi_i(x_i) = 1$. We will denote the finite element approximation to a scalar function v by an upper-case letter V and the corresponding vector of nodal values by a boldface character $\mathbf{V} = (\mathbf{V}_i)_{i \in I_h}$ such that $V = \sum_{i \in I_h} \mathbf{V}_i \varphi_i$. Similarly, a discretised deformation ϕ is expressed as $\Phi = \sum_{i \in I_h} \sum_{j=1}^d \Phi_{ij} \varphi_i e_j$, where e_1, \dots, e_d denotes the canonical Euclidean basis of \mathbb{R}^d .

For simplicity, let us here only consider the pure averaging model without simultaneous image segmentation. Using the above notation, the discretised Euler–Lagrange equations from Section 4.2.2 read

$$\begin{aligned} 0 &= \left(\frac{2\gamma}{n\varepsilon} \sum_{i=1}^n M \left[((1 - U_i \circ \Phi_i^{-1})^2 + (U_i \circ \Phi_i^{-1})^2) |\det(\mathcal{D}\Phi_i^{-1})| \right] + \eta\varepsilon L[1] + \frac{\eta}{\varepsilon} M[1] \right) \mathbf{U} \\ &\quad - \left(\frac{2\gamma}{n\varepsilon} \sum_{i=1}^n M \left[(U_i \circ \Phi_i^{-1})^2 |\det(\mathcal{D}\Phi_i^{-1})| \right] + \frac{\eta}{\varepsilon} M[1] \right) \mathbf{1} =: A_{U_i, \Phi_i} \mathbf{U} - b_{U_i, \Phi_i}, \\ 0 &= \frac{2\gamma}{n\varepsilon} \int_{\Omega} \left((1 - U_i)^2 (U \circ \Phi_i) - U_i^2 (1 - U \circ \Phi_i) \right) \Psi \cdot (\nabla U \circ \Phi_i) \, dx \\ &\quad + \frac{1}{n} \int_{\Omega} \left((1 - \delta) \chi_{\mathcal{O}_i}^\varepsilon + \delta \right) W_{,A}(\mathcal{D}\Phi_i) : \mathcal{D}\Psi \, dx, \end{aligned}$$

where for some weighting function $\omega : \Omega \rightarrow \mathbb{R}$ we have defined the generalised mass matrix $M[\omega]$ and the generalised stiffness matrix $L[\omega]$ as

$$M[\omega] = \left(\int_{\Omega} \omega \varphi_i \varphi_j \, dx \right)_{ij}, \quad L[\omega] = \left(\int_{\Omega} \omega \nabla \varphi_i \cdot \nabla \varphi_j \, dx \right)_{ij},$$

where $\mathbf{1}$ is the nodal vector with entries all equal to 1, and where the vector-valued test function Ψ runs over all basis functions $\varphi_i e_j$. Overall, this represents $(nd+1)N$ nonlinear equations in the unknowns $\mathbf{U} \in \mathbb{R}^N$ and $\Phi_1, \dots, \Phi_n \in \mathbb{R}^{dN}$, where $N = (2^L + 1)^d$.

The integrals in the above equations are evaluated using Gaussian quadrature of third order on each grid cell. Pullbacks $U \circ \Phi$ are computed exactly at the quadrature points, whereas pushforwards $U_i \circ \Phi^{-1}$ are approximated as $U_i \circ \mathcal{I}_h(\Phi^{-1})$ for the nodal interpolation operator \mathcal{I}_h . To obtain this nodal interpolation of the inverse deformation we proceed as follows. We map each grid cell under the deformation Φ onto the image domain and identify all grid nodes which are located within this deformed cell. Next, we need to find those points inside the grid cell for which the local, multilinear interpolation of Φ retrieves these grid nodes. This yields a small system of nonlinear equations which

can be solved by few Newton iteration steps. Let us emphasise that it is not sufficient just to compute the nodal interpolants $\mathcal{I}_h(U \circ \Phi_i)$, $\mathcal{I}_h(U_i \circ \Phi_i^{-1})$ of $U \circ \Phi_i$ and $U_i \circ \Phi_i^{-1}$, respectively. Indeed, artificial displacements near the shape edges are then observed, accompanied by strong tensions and generated while alternating between optimising the average phase field U and the deformations Φ_i .

In order to solve the above equations we employ a gradient descent type scheme: Note that the discrete Euler–Lagrange equation in the phase field vector \mathbf{U} is linear and hence can be solved directly using a conjugate gradient iteration. Hence, in each step of our algorithm we will first solve for the average phase field and then perform $k \geq 1$ regularised gradient descent steps for the deformations Φ_i according to

$$\Phi_i = \Phi_i^{\text{old}} - \tau \text{grad}_{\Phi_i} \mathcal{E}^{\gamma, \varepsilon},$$

where τ is the current step size and grad_{Φ_i} represents the finite element approximation to the gradient with respect to the weighted $H^1(\Omega)$ inner product

$$(\Psi_1, \Psi_2)_\sigma := (\Psi_1, \Psi_2)_{L^2(\Omega)} + \frac{\sigma^2}{2} (\mathcal{D}\Psi_1, \mathcal{D}\Psi_2)_{L^2(\Omega)}.$$

More precisely, $\text{grad}_{\Phi_i} \mathcal{E}^{\gamma, \varepsilon}$ is the vector-valued finite element function defined as the solution of $(\text{grad}_{\Phi_i} \mathcal{E}^{\gamma, \varepsilon}, \Psi)_\sigma = \langle \delta_{\Phi_i} \mathcal{E}^{\gamma, \varepsilon}, \Psi \rangle$ for all discrete displacement fields Ψ (where the right-hand side of this equation coincides with the right-hand side of the Euler–Lagrange equation). The effect of the smoothing metric $(\cdot, \cdot)_\sigma$ is related to the convolution with a Gauss kernel or equivalently the application of one time step for the heat equation semigroup. Consequently, information flow across the image is enhanced, and the deformations equilibrate faster. Also, the descent algorithm becomes more resistant to being trapped in local minima. In the algorithm it turns out to be sufficient to approximate the gradient performing a single multigrid V-cycle for the system

$$\left(M[1] + \frac{\sigma^2}{2} L[1] \right) (\text{grad}_{\Phi_i} \mathcal{E}^{\gamma, \varepsilon})_j = (\langle \delta_{\Phi_i} \mathcal{E}^{\gamma, \varepsilon}, \Psi \rangle)_j,$$

where $(\cdot)_j$ for $j = 1, \dots, d$ indicates one particular component sub-vector of the nodal vector and—with a slight misuse of notation— $\text{grad}_{\Phi_i} \mathcal{E}^{\gamma, \varepsilon}$ and $\langle \delta_{\Phi_i} \mathcal{E}^{\gamma, \varepsilon}, \Psi \rangle$ are interpreted as nodal vectors (see [39, 50] for details).

The step size τ is obtained according to Armijo’s rule, which ensures sufficient agreement between the objective functional and its linearisation. If the actually observed energy decay in one time step is smaller than $\frac{1}{4}$ of the decay estimated from the gradient (the Armijo condition is then violated), then the time step τ is halved for the next trial, else it is doubled as often as possible without violating the Armijo condition.

In order to enable a fast energy relaxation as well as to avoid local minima we try to obtain the large-scale part of the deformations ϕ_i first by minimising the energy $\mathcal{E}^{\gamma, \varepsilon}$ on a coarse spatial resolution, that is, a grid on $[0, 1]^d$ with just $2^{l_0} + 1$ vertices in each space direction for $l_0 < L$. Afterwards, the obtained results are prolonged via multilinear

interpolation onto the next grid level $l_0 + 1$ with $2^{(l_0+1)} + 1$ nodes in each direction, and the minimisation is continued at this finer resolution. The procedure is iterated until the finest grid level L . In fact, this approach is the reason for choosing a dyadic grid resolution, which strongly simplifies the implementation of a grid hierarchy for a multi-scale algorithm. Since the width ε of the diffusive phase field edges should naturally scale with the grid width h , we choose $\varepsilon = Ch$ on each level for some constant factor C (which is set to one in the implementation).

The entire algorithm in pseudo code notation reads as follows:

```

EnergyRelaxation  $((Y_i^0)_{i=1,\dots,n})$  {
  initialise  $\Phi_i = \text{id}$  on grid level  $l_0$  for all  $i = 1, \dots, n$ ;
  for grid level  $l = l_0$  to  $L$  {
    do {
      segment the images  $(Y_i^0)_{i=1,\dots,n}$  to obtain phase fields  $(U_i)_{i=1,\dots,n}$ ;
       $\mathbf{U}^{\text{old}} = \mathbf{U}$ ;
      solve the linear system
         $A_{U_i, \Phi_i} \mathbf{U} = b_{U_i, \Phi_i}$ 
        for the phase field vector  $\mathbf{U}$ ;
      for image  $i = 1$  to  $n$ 
        for count  $k = 1$  to  $K$  {
           $\Phi_i^{\text{old}} = \Phi_i$ ;
          perform a gradient descent step
             $\Phi_i = \Phi_i^{\text{old}} - \tau \text{grad}_{\Phi_i^{\text{old}}} \mathcal{E}^{\gamma, \varepsilon}[\mathbf{U}, (\Phi_j)_{j=1,\dots,n}]$ 
            with Armijo step size control for  $\tau$ ;
        }
      } while  $(\sum_{i=1}^n |\Phi_i^{\text{old}} - \Phi_i| + |\mathbf{U}^{\text{old}} - \mathbf{U}| \geq \text{Threshold})$ ;
      if  $(l < L)$  prolongate  $\mathbf{U}, \Phi_i$  for all  $i = 1, \dots, n$  onto the next grid level;
    }
  }
}

```

In the computations, values of $l_0 = 4$, $K = 5$, and $\text{Threshold} = 2 \cdot 10^{-4}$ seemed to yield appropriate progress during the energy minimisation. For the example considered in Figure 4.2, one iteration on grid level $L = 9$ takes 10 seconds on a Pentium IV PC at 1.8 GHz, running under Linux. The complete method typically converges after roughly 100 such iterations on each grid level.

In Figure 4.9, we depict the progression of the various energy components of $\mathcal{E}^{\gamma, \varepsilon}$ for the averaging problem from Figure 4.2. The strong decay of the global energy at the beginning of the algorithm is clearly visible. Apparently, the mismatch penalty strongly dominates the total energy. Also, we show the L^1 -difference between $u_1 \circ \phi_1^{-1}$ and $u_2 \circ \phi_2^{-1}$, which strongly decreases, indicating a good match between both deformed shapes.

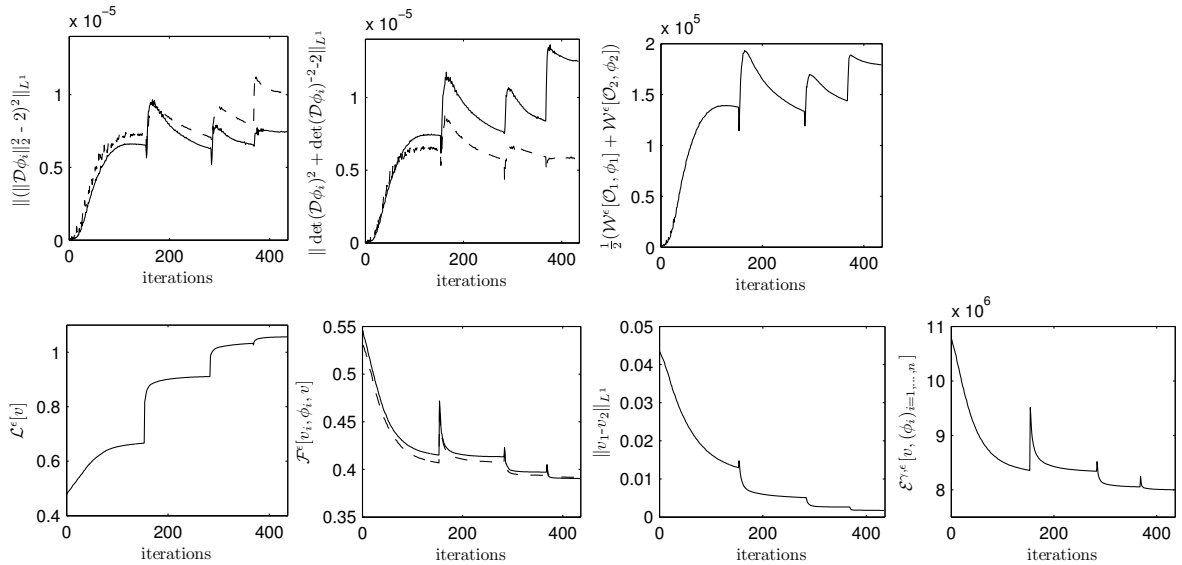


Figure 4.9: The progression of the various energy contributions during the algorithm is shown for the example from Figure 4.2. The top row shows the hyperelastic energy contributions due to length and volume variation (left and middle, respectively; the solid line corresponds to $i = 1$, the dashed one to $i = 2$) as well as the total hyperelastic energy (right), that is, the sum of length and volume contributions. The bottom row shows the length regularisation (left) and the mismatch penalty for both images (second graph), as well as the overall energy (right). In all graphs, the spikes correspond to the prolongation to the next grid level. The third graph in the bottom row shows the L^1 -difference between $u_1 \circ \phi_1^{-1}$ and $u_2 \circ \phi_2^{-1}$.

4.4 Validation and experiments

In the following, the shape averaging approach is applied to various collections of 2D and 3D shapes and to image morphologies.

4.4.1 Averaging of 2D shapes

As first illustrative examples, the average of different 2D objects is shown in Figures 4.10 to 4.11. Furthermore, Figures 4.2 and 4.3 have already shown that, due to the invariance of the hyperelastic energy with respect to local rotations, the computed averages try to locally preserve isometries. Effectively, the different characteristics of the input shapes, both on the global and a local scale, are averaged in a physically intuitive way, and the scheme proves to be fairly robust due to the diffusive approximation based on the phase field model and the multi-scale relaxation. Nevertheless, a lack of topological

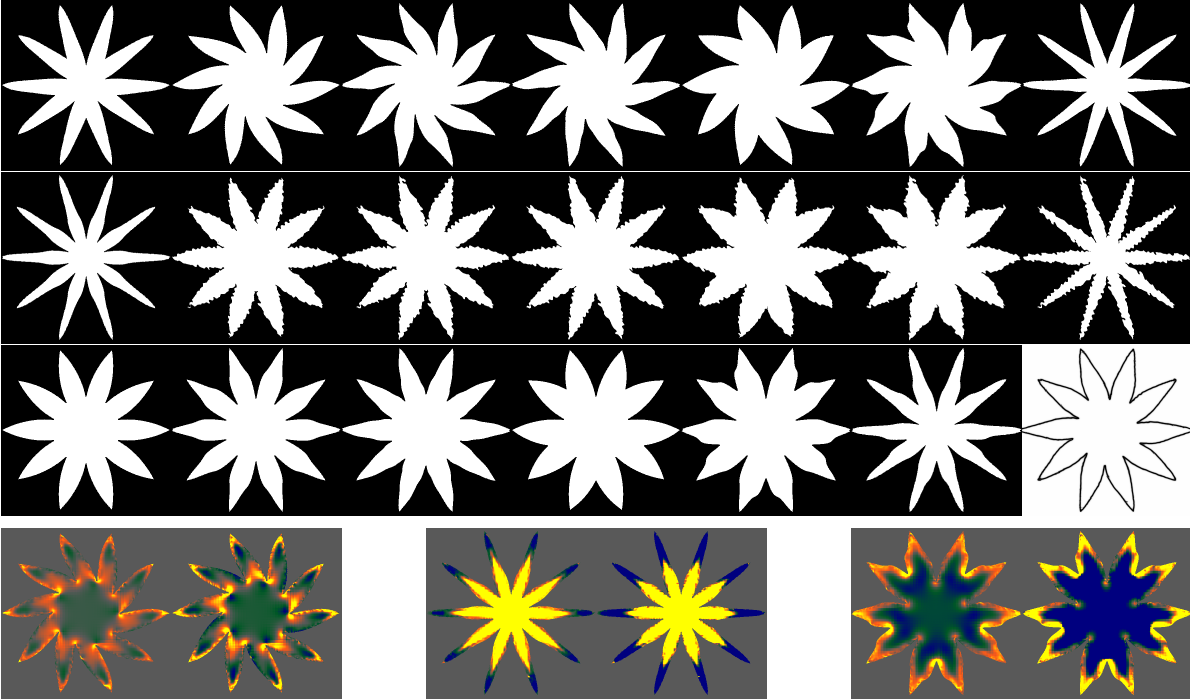



Figure 4.10: 20 shapes “device7” from the MPEG7 shape database and their average phase field. The bottom line shows $\frac{1}{\sqrt{2}}\|\mathcal{D}\phi_i\|_F$ and $\det(\mathcal{D}\phi_i)$ for shape 2, 8, and 19, with ranges of $[0.8, 1.2]$ colour-coded as . (Resolution 513^2 , $(a, b, \gamma, \eta) = (0.1, 0.1, 1, 10^{-9})$)

equivalence of the given input shapes might lead to corresponding local artifacts in the shape average result, compare for instance the twentieth input shape in Figure 4.11 and its locally spurious impact on the average phase field between the two legs.

Figure 4.12 shows two more examples, using input shapes from [43] and the shape database at the Centre for Vision, Speech, and Signal Processing, University of Surrey. Averaging the hand shapes yields very similar results to the averages obtained in [43] and [59] as the Euclidean and the Fréchet mean of vectors of landmark positions, respectively. The average fish shape has also been computed in [31]. Note that our result preserves more fine structures as opposed to the quite rounded mean shape in [31].

4.4.2 Averaging of 3D shapes

In what follows we consider the averaging of 3D shapes originally given as triangulated surfaces and first converted to an implicit representation as binary images. A set of 48 kidneys and a set of 24 feet will serve as input data. The first five original kidneys and their computed average have already been shown in Figure 4.7. Local structures seem to be quite well represented and preserved during the averaging process compared to for

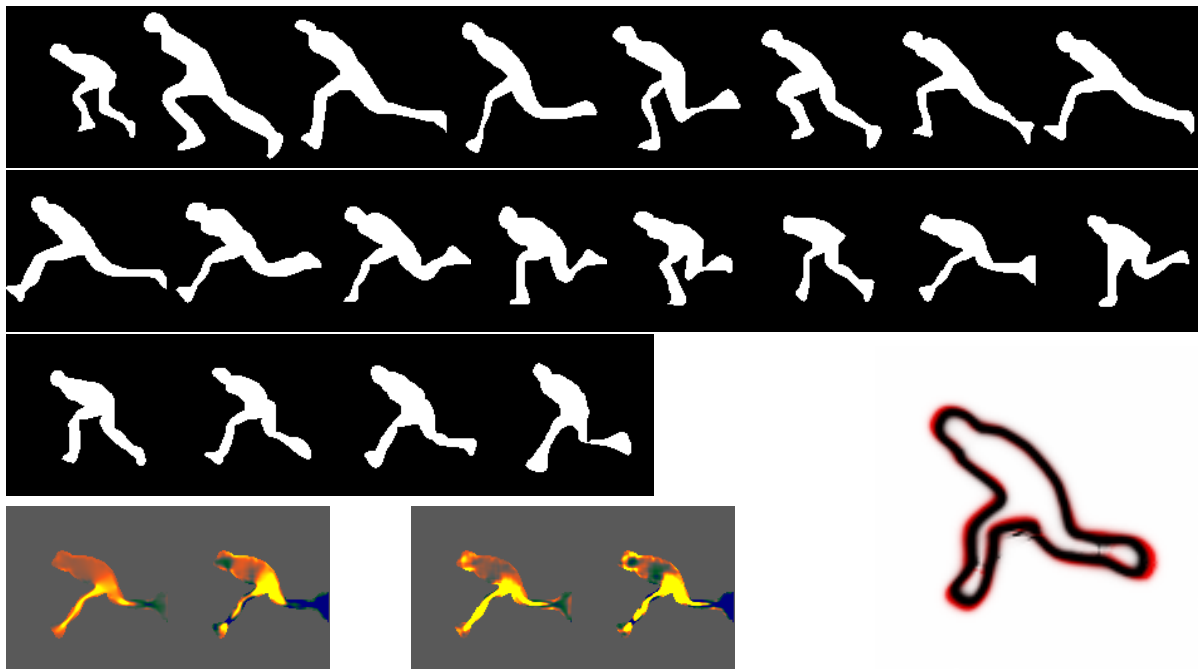


Figure 4.11: 20 shapes “stef” from the MPEG7 shape database and their average phase field (bottom right) for length change penalisation ten times as large as volume change penalisation $((a, b) = (1, 0.1))$, black, on top) and the other way round $((a, b) = (0.01, 0.1))$, red, underneath). The bottom line shows $\frac{1}{\sqrt{2}}\|\mathcal{D}\phi_i\|_F$ and $\det(\mathcal{D}\phi_i)$ for shape 15 in both cases with ranges of $[0.8, 1.2]$ colour-coded as . Obviously, the larger the ratio between the weights of volume and length variation penalty, the more elongated the shapes become. (Resolution 129^2 , $(\gamma, \eta) = (1, 10^{-9})$)

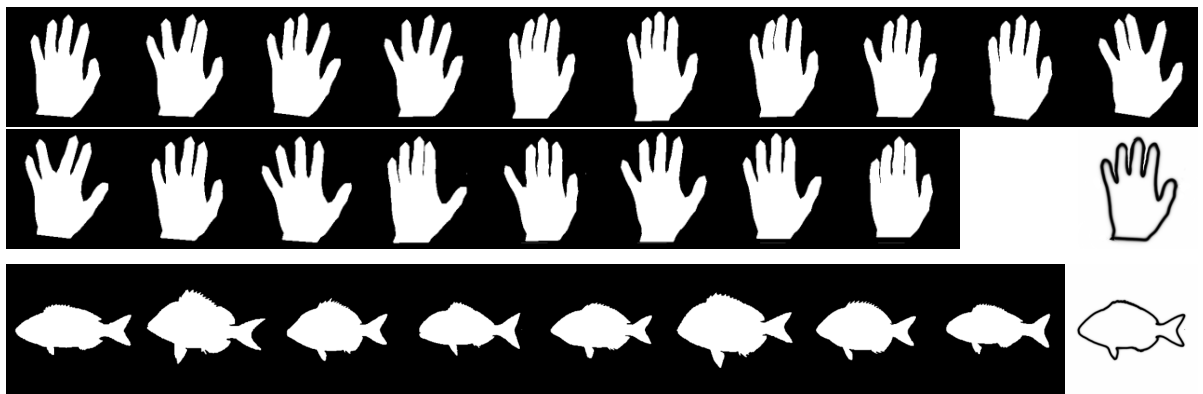


Figure 4.12: Average of 18 hand and 8 fish silhouettes, taken from [43] and the shape database at the Centre for Vision, Speech, and Signal Processing, University of Surrey, respectively.

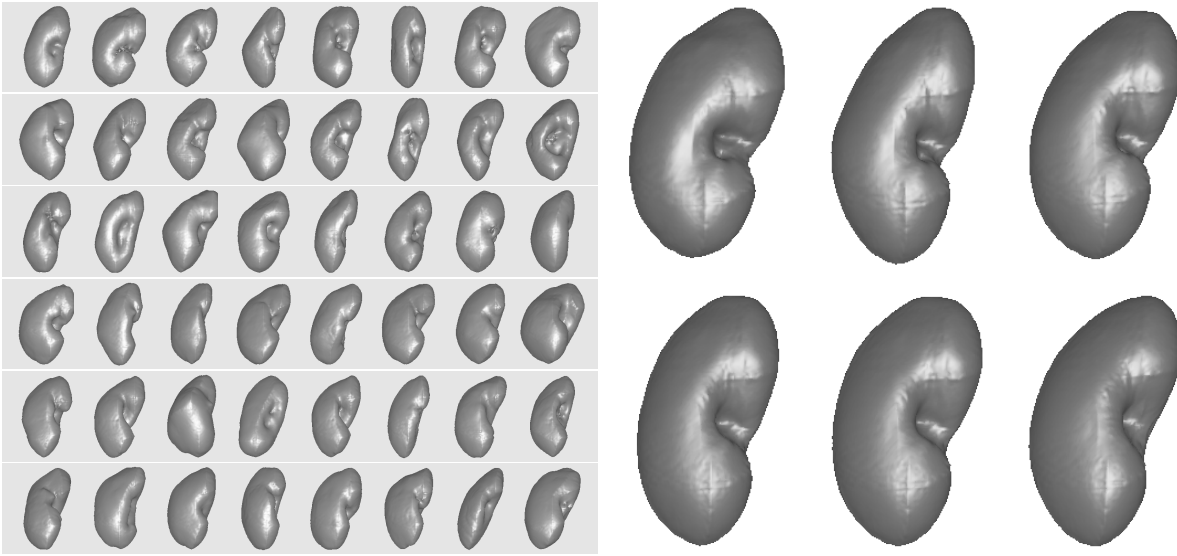


Figure 4.13: On the left, 48 kidney shapes are shown. On the right, from top left to bottom right the average shape of the first two, four, five, six, eight, and of all 48 kidneys is depicted. (Resolution 257^3 , $(a, b, \gamma, \eta) = (10, 1, 1, 10^{-7})$)

example the average of kidney shapes in medial representation in [57].

Via the deformation of a given object onto the average during the averaging algorithm, the method also yields local and global information about the distance of the object to the average. For the first two of the original kidneys, the corresponding local elastic energy density or rather the three deformation invariants are also depicted in Figure 4.7. The inside of the first kidney apparently gets slightly dilated, whereas the second one is compressed. Also, it can be observed that the dilation or compression is reduced at the shape boundaries, which is caused by the finite width phase field description of the edges: For the deformed phase fields to match, they all have to have the same thickness and hence may only be deformed significantly in tangential, but not in normal direction.

Naturally, any averaging will involve some smoothing, eliminating fine details which differ from shape to shape. It is hence of interest whether features, common to all shapes but differing slightly, pertain if the number of samples is increased: Figure 4.13 shows the result of averaging different numbers of kidneys. Also, we would like to know how the method performs for relatively large numbers of input shapes, since the lack of a triangle inequality for a hyperelastic “distance” measure raises the question whether a law of large numbers holds. It is indeed observed that the middle dent of the average kidney is a little less pronounced than in each single kidney, even in the case of averaging just two kidneys. Also, the influence of each additional original shape seems significantly strong, however, a kidney-like shape is doubtlessly preserved up to the average of all 48 kidneys.

The next example consists of a set of feet, where the average may help to design an

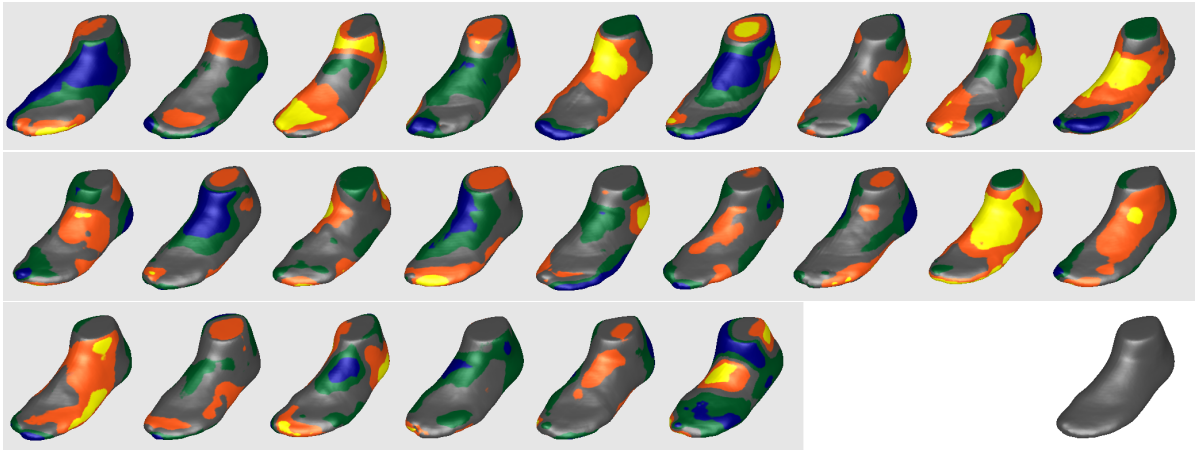


Figure 4.14: 24 given foot shapes, textured with the distance to the surface of the average foot (bottom right). Values range from 6 mm inside the average foot to 6 mm outside, colour-coded as . The front of the instep can be identified as a region of comparatively low variation. (Resolution 257^3 , $(a, b, \gamma, \eta) = (10, 10, 1, 10^{-7})$)

optimal shoe. The 24 original feet are displayed in Figure 4.14. Their surface is coloured according to the local distance to the surface of the computed average shape, which helps to identify regions of strong variation. For that purpose, the foot shapes have been optimally aligned with the average for the final visualisation (while the algorithm itself robustly deals with even quite large rigid body motions). Apparently, the instep differs comparatively little between the given feet, whereas the toes show a rather strong variation. Note that—since we only display normal distance to the surface of the average foot—any potential tangential displacement is not visible, but could of course also be visualised when examining shape variation.

For real applications, one has to be careful when dealing with shapes of different volume. Depending on the chosen hyperelastic parameters, there may be a bias towards larger or smaller shapes, and appropriate parameters will have to be chosen carefully. Also, for too soft hyperelastic material models, buckling instabilities will occur during compression of large volume shapes. Some influence of different hyperelastic parameters is illustrated in Figure 4.11.

4.4.3 Weighted averaging

Returning back to the kidneys, it is also possible to compute a weighted average, where the deformation energy of the different input shapes is weighted differently. The aver-

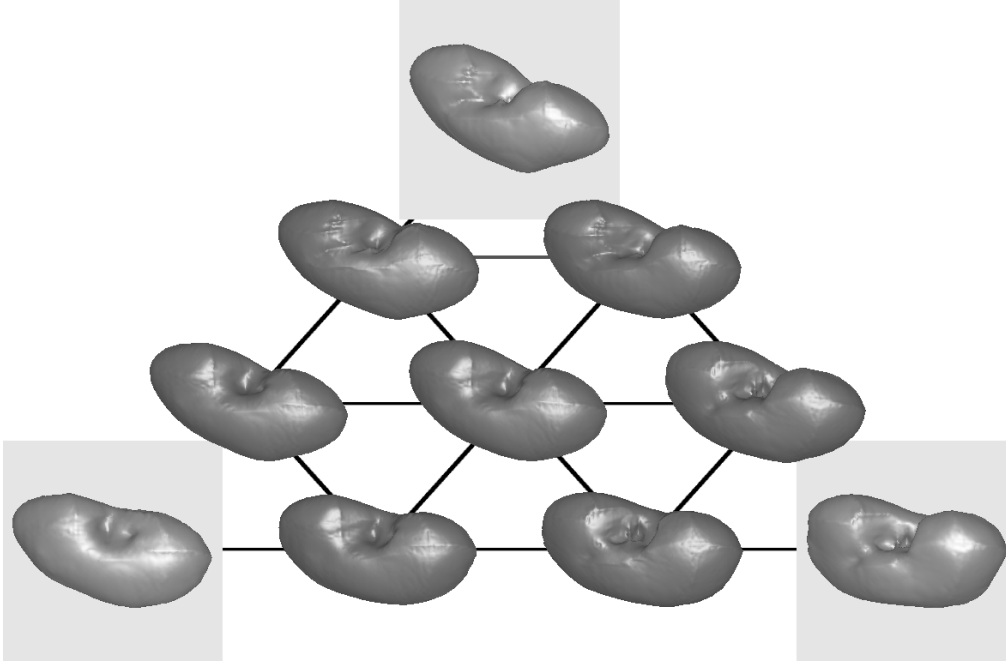


Figure 4.15: Given the three kidney geometries placed in the corners of the triangle, seven differently weighted averages are placed at corresponding positions in the triangle.

aging functional then is modified to

$$\mathcal{E}^{\gamma, \varepsilon}[u, (\phi_i)_{i=1, \dots, n}] = \sum_{i=1}^n \left(\lambda_i \mathcal{W}^\varepsilon[\mathcal{O}_i, \phi_i] + \frac{\gamma}{n} \mathcal{F}^\varepsilon[u_i, \phi_i, u] \right) + \eta \mathcal{L}_{\text{AT}}^\varepsilon[u],$$

where the weights λ_i are nonnegative and add up to 1. Such a “nonlinear convex-combination” of three kidneys is presented in Figure 4.15.

4.4.4 Averaging image morphologies

To illustrate that the approach can also be applied to average image morphologies, the input of the final example consists of two-dimensional, transversal CT scans of the human thorax from four different patients (Figure 4.16, left). Unlike the previous examples, these images do not encode volumetric shapes homeomorphic to the unit ball, but contain far more complicated structures. Also, the quality of contrast differs between the images, and—even more problematic—the images do not show a one-to-one correspondence, that is, several structures (the scapula, ribs, parts of the liver) are only visible in some images, but not in others, implying that the underlying shapes are not even homeomorphic. Nevertheless, the algorithm manages to segment and align the main features (the heart, the spine, the aorta, the sternum, the ribs, the back muscles, the

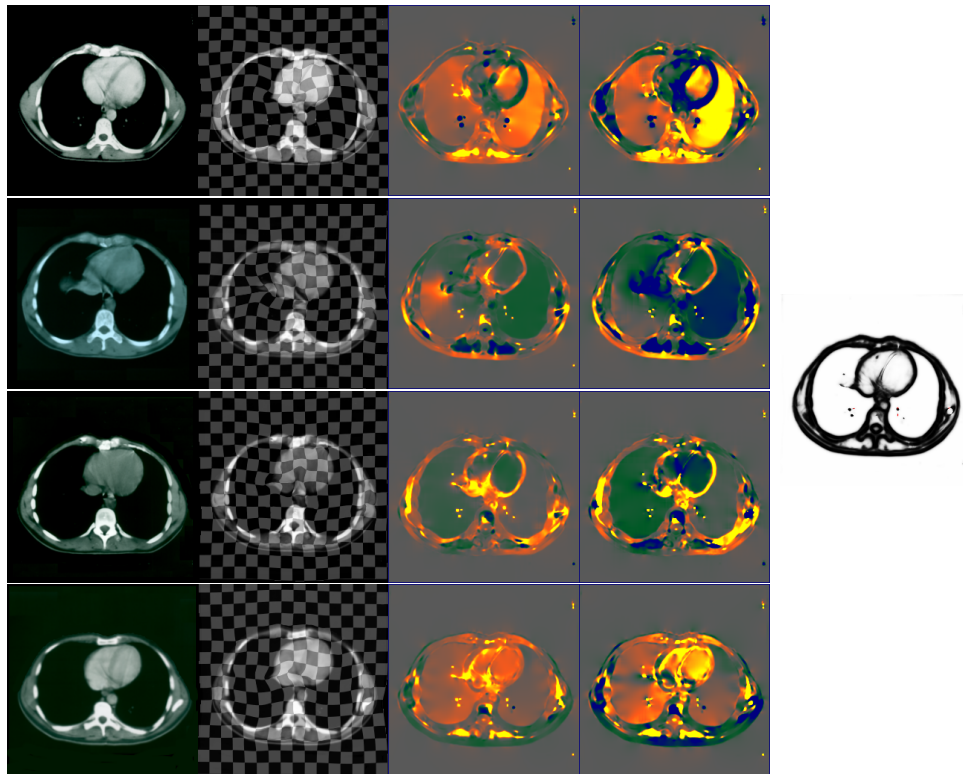



Figure 4.16: Averaging CT scan slices of the thorax from four different patients. From left to right: Original images, deformations ϕ_i (applied to a chequerboard on which the original image was printed), $\frac{1}{\sqrt{2}}\|\mathcal{D}\phi_i\|_F$ and $\det(\mathcal{D}\phi_i)$ (colour-coded as  with range $[0.8, 1.2]$), and average phase field.

skin), yielding sensible average contours (Figure 4.16, right). In order to achieve this, we this time jointly segmented and averaged the original CT scans using the joint model $\mathcal{E}_{\text{joint}}^{\gamma, \varepsilon}$. The second to fourth column of Figure 4.16 depict the corresponding deformations ϕ_i and the deformation invariants. Obviously, the deformation behaves quite regularly: Not only is it homeomorphic, but also too large and distorting deformations are prevented by the hyperelastic regularisation. This enables the method to be applied to images containing also distinct structures, whereas for viscous flow regularisation as in [19, 36] such individual structures are at risk of being matched with anything nearby. The deformation energy is quite evenly distributed over the images and only peaks at pronounced features, where a local exact fit can be achieved (for instance, at the back muscles). Outside the thorax, the energy rapidly decreases to zero, justifying that in this example we did not weight the elastic energy differently inside and outside the body.

4.5 Discussion

The proposed averaging concept establishes a connection between a purely geometric and a physically motivated view on shapes: From the geometric viewpoint, the distance between two shapes or objects is supposed to indicate how far these are from being isometric. Hyperelasticity is here employed as a means to retrieve this information on a physical basis. The use of a nonlinear elastic energy is of crucial importance so as to be capable of measuring distortions correctly despite superposition of local rotations. Also the nonlinearity inside the material law can be exploited to assign varying importance to the volumetric, area-distorting, and length-changing components of the distortion.

Having computed averages of shapes, the natural next step is to consider the higher moments of shape variation and, in particular, to analyse the directions of strongest shape variation within a given sample of shapes (which is nontrivial since the elastic shape space neither exhibits the structure of a linear vector space nor of a Riemannian manifold, compare Section 1.1). This subject will be addressed in the next chapter by determining some kind of eigenfunctions of an appropriate covariance operator. The resulting dominant modes of variation could then for example be used to obtain a low-dimensional shape representation or some Mahalanobis distance which induces a shape prior.

In the proposed variational approach, it was possible to integrate the actual cost functional (in this case the accumulated elastic deformation energy) and the constraints in one single objective functional. As constraints we may in fact not only consider the congruence between the deformed input shapes, here implemented as a mismatch penalty, but also the fact that the matching deformations should obey the principle of momentum conservation, that is, that the deformations are physically sensible in that the induced stress field is divergence-free and that the boundary stresses of the different input shapes balance. Fortunately, this is equivalent to the deformations being minimisers of the elastic deformation energy which was exactly our primal objective functional. However, one might want to choose a different objective functional for averaging or in the more general context of shape optimisation, which for example depends on the boundary stresses induced by the deformation. Then, the constraint that the deformations have to be minimisers of the elastic deformation energy cannot be incorporated in the objective functional, but has to be dealt with separately. An optimisation problem with an additional constraint of this type is treated in Chapter 7.

The Ambrosio–Tortorelli phase field has the nice property of encoding just edge sets which need not necessarily be the boundary of some object so that the averaging method can also be applied to averaging image morphologies. Also, the Ambrosio–Tortorelli regularisation is quadratic, which allows for a direct and fast solution for the phase field. However, the Ambrosio–Tortorelli phase field is not generic to describe objects with a volume, for which case it would be nice to have an intrinsic distinction between the inside of an object and its outside. For this reason we will employ a level set description of shapes in Chapter 6 and Modica–Mortola phase fields in Chapter 7.

As a final remark, the elastic approach does not allow for a canonic definition of paths in shape space, which however is desired in some cases, for example in order to do shape warping in computer vision. For such problems it is preferable to equip the shape space with a Riemannian structure which will be done in Chapter 6.

5 Principal modes of elastic shape variation

The following chapter is devoted to the analysis of shape variations based on the model of shapes as boundaries of elastic objects as introduced in the previous chapter. This may be seen as the natural next step after having computed a shape average. Applications of such a second moment analysis are manifold; the results can, for example, be used as shape priors in image segmentation or object detection.

We will introduce a mechanically sound linearisation of shape variations which will enable the definition of a physically meaningful covariance operator that can be used for a principal component analysis. Unlike in a Riemannian setting, this linearisation cannot be based on an intrinsic notion of paths in shape space. In fact, on a Riemannian manifold, the Riemannian metric induces the concept of geodesic paths between any two points $\mathcal{S}_1, \mathcal{S}_2$ on the manifold, whose length provides a distance $d(\mathcal{S}_1, \mathcal{S}_2)$ between them. An average of given points $\mathcal{S}_1, \dots, \mathcal{S}_n$ can then be computed as the minimiser of the sum of squared distances,

$$\mathcal{S} = \arg \min_{\tilde{\mathcal{S}}} \sum_{i=1}^n d(\mathcal{S}_i, \tilde{\mathcal{S}})^2.$$

The logarithmic map at the average \mathcal{S} then associates each input point \mathcal{S}_i with a vector element $\log_{\mathcal{S}} \mathcal{S}_i$ of the tangent space to the manifold at \mathcal{S} . This vector $\log_{\mathcal{S}} \mathcal{S}_i$ represents the initial velocity of the geodesic connecting \mathcal{S} with \mathcal{S}_i and satisfies $(\log_{\mathcal{S}} \mathcal{S}_i)^2 = d(\mathcal{S}, \mathcal{S}_i)^2$. For sure, this framework can only be applied in regions of the manifold where geodesics are unique and thus the logarithm is well-defined (we will encounter a similar limitation in our elastic setup). Otherwise there might for example be conjugate points that are connected by a continuous family of geodesics such as the north and the south pole on the two-dimensional sphere for which an average obviously cannot be defined. Nevertheless, the logarithms may be seen as linear representatives of the input points, and they satisfy

$$0 = \sum_{i=1}^n \log_{\mathcal{S}} \mathcal{S}_i.$$

Since they point into the initial direction of the geodesic which leads to \mathcal{S}_i , they are in this sense linearisations of the variation of the average \mathcal{S} in the direction of the different input points, on which a principal component analysis can readily be performed. In the shape space with an elastic structure, however, the fundamental axiom of elasticity

prevents a straightforward definition of geodesic paths: When deforming a shape \mathcal{S}_1 into another one \mathcal{S}_2 , then the final state, the stress configuration, and the needed deformation energy are independent of the path along which \mathcal{S}_1 was deformed; all possible paths that yield the same final deformation are in this sense equal.

We will employ the following conceptual idea. The first moment analysis from the previous chapter already yields an average shape $\mathcal{S} = \partial\mathcal{O}$ as well as matching deformations ϕ_i that deform the input objects \mathcal{O}_i and shapes $\mathcal{S}_i = \partial\mathcal{O}_i$ into \mathcal{O} and $\mathcal{S} = \partial\mathcal{O}$, respectively. These deformations induce boundary stresses $\sigma_i\nu$ on \mathcal{S} . An infinitesimal modulation of these stresses results in displacements of the average shape \mathcal{S} , which may be seen as linearisations of the variation of the average \mathcal{S} in direction of the input shapes. A principal component analysis can then be performed on these displacements. In this approach, care has to be taken to adequately take into account the particular nature of the average object \mathcal{O} as a composition of deformed (and thus prestressed) input objects \mathcal{O}_i . Note that we here properly describe truly nonlinear variations due to the linearisation via stresses which are induced by the nonlinear deformations.

Via the spring analogy from the previous chapter, this procedure can still be related to the statistical analysis of a set of points $x_1, \dots, x_n \in \mathbb{R}^d$. Recall that the average x can be represented as the minimiser of the total elastic spring energy for a set of elastic springs connecting x_i , $i = 1, \dots, n$, with x . While—due to Hooke’s law—the energy of each spring is given by $\frac{D}{2}|x_i - x|^2$ for the spring constant D , the force exerted on x is $D(x_i - x)$. If a small fraction $\delta D(x_j - x)$ of the j th spring force is added, the point x gets displaced by $\frac{\delta}{n}(x_j - x)$. Hence, the covariance tensor $((x_i - x) \cdot (x_j - x))_{i,j=1,\dots,n}$ of the input points can—up to a multiplicative constant—be identified with the covariance tensor of these displacements.

As in the previous chapter, it is in principle possible to base the statistical analysis on a variety of nonlinear elasticity models. Furthermore, we will be able to choose the metric in the definition of the covariance operator in different ways: A standard L^2 -metric pronounces shape variations with large displacements even though they are energetically cheap (for example, a rotation of some structure around a joint), while the Hessian of the nonlinear elastic energy serves as the appropriate inner product so as to measure distances between displacements solely based on the associated change of elastic energy. Indeed, the Hessian represents an averaged linearised elasticity tensor at the deformed configuration. Thus, displacements in regions and directions which are significantly loaded are weighted strongly, which is mechanically sound.

In the following, we will first introduce the linearisation of shape variations in Section 5.1 and then present the corresponding covariance analysis in Section 5.2. Afterwards, we will describe the numerical implementation in Section 5.3 and finally show some applications in Section 5.4.

5.1 Linearisation of shape variations

As mentioned earlier, the averaging procedure from the previous chapter provides us not only with an average shape $\mathcal{S} = \partial\mathcal{O}$ of input shapes $\mathcal{S}_i = \partial\mathcal{O}_i$, $i = 1, \dots, n$, but also with deformations ϕ_i such that $\phi_i(\mathcal{O}_i) = \mathcal{O}$. These deformations ϕ_i induce Cauchy boundary stresses $\sigma_i\nu[\mathcal{S}]$ on the average shape \mathcal{S} or equivalently first Piola–Kirchhoff stresses $\sigma_i^{\text{ref}}\nu[\mathcal{S}_i]$ on the original shape \mathcal{S}_i (compare Section 3.1), where $\nu[\mathcal{S}]$ and $\nu[\mathcal{S}_i]$ denote the unit outward normals on \mathcal{S} and \mathcal{S}_i , respectively. The boundary stresses of one deformation ϕ_i are counteracted by the stresses from the deformations of the other input shapes (see Figure 4.5), otherwise the deformed shape $\phi_i(\mathcal{S}_i)$ would snap back to its original configuration.

Let us assume that the relation between boundary stresses and deformations is one-to-one. If the boundary stress $\sigma_i^{\text{ref}}\nu[\mathcal{S}_i]$ is scaled by a parameter $\delta \in [0, 1]$, this new boundary stress corresponds to a deformation different from ϕ_i . In fact, the corresponding deformation will be exactly that one which results from the surface load $\delta\sigma_i^{\text{ref}}\nu[\mathcal{S}_i]$ being applied to the boundary of object \mathcal{O}_i . The scaling of the boundary stress hence produces a one-parameter family of deformations, connecting a vanishing deformation (for $\delta = 0$) with ϕ_i (for $\delta = 1$). In this sense, the scaled boundary stress generates a path in shape space between the original input shape \mathcal{S}_i and the average \mathcal{S} , and the stress may be seen as a linear representative of input shape \mathcal{S}_i .

Similarly to the Riemannian approach, we can now examine the variation of the average \mathcal{S} as the impact of input shape \mathcal{S}_i is increased. For this purpose, we apply a small fraction δn of the Cauchy boundary stress $\sigma_i\nu[\mathcal{S}]$ at the average shape boundary \mathcal{S} , thereby inducing a small displacement δv_i of all points in the average object \mathcal{O} . (The scaling by n just serves to divide out the dependence on the number of input shapes.) It is these v_i that we will later employ to perform a covariance analysis; they represent the desired linearisations of shape variations among $\mathcal{S}_1, \dots, \mathcal{S}_n$. Here, \mathcal{O} should be interpreted as the composition of all deformed and thus prestressed input objects $\phi_i(\mathcal{O}_i)$; an elasticity model that assumes \mathcal{O} to be internally relaxed would not take into account the nonlinear structure. In detail, the scaled Cauchy stress $\delta n\sigma_i\nu[\mathcal{S}]$ of object \mathcal{O}_i here acts as first Piola–Kirchhoff stress on the compound object \mathcal{O} . For small δ , this is equivalent to performing a weighted average as in Section 4.4.3, where \mathcal{S}_i is assigned the weight $\frac{1+n\delta}{n(1+\delta)}$ while all other shapes are weighted with $\frac{1}{n(1+\delta)}$.

5.1.1 The associated mechanical problem

To properly model the loaded configurations, we concatenate each deformation ϕ_k with the collective deformation $\text{id} + \delta v_i$. The equilibrium displacements δv_i can then be obtained by minimising the nonlinear energy

$$\mathcal{E}_i^\delta[v_i] = \frac{1}{n} \sum_{k=1}^n \mathcal{W}[\mathcal{O}_k, (\text{id} + \delta v_i) \circ \phi_k] - \delta^2 \int_{\mathcal{S}} \sigma_i\nu[\mathcal{S}] \cdot v_i \, da$$

(as explained in Section 3.1), where \mathcal{W} denotes the hyperelastic deformation energy from the previous chapter and $\delta n\sigma_i\nu[\mathcal{S}]$ acts as a surface load. Upon integration by parts and using the fact that $\operatorname{div} \sigma_i = 0$ holds on \mathcal{O} (see Section 4.1.1), the boundary integral can also be replaced by the volume integral $\int_{\mathcal{O}} \sigma_i : \mathcal{D}v_i \, dx$, which is more convenient with respect to a numerical discretisation. Since this variational definition determines v_i only up to rigid body motions, we have to impose the additional constraints of zero average displacement and angular momentum,

$$\int_{\mathcal{O}} v_i \, dx = 0 \quad \text{and} \quad \int_{\mathcal{O}} x \times v_i \, dx = 0.$$

The minimising δv_i indeed describes the displacement induced by applying the additional boundary stress $\delta n\sigma_i\nu[\mathcal{S}]$: Abbreviating $\psi_i := \operatorname{id} + \delta v_i$, the Euler–Lagrange equation for the minimisation of $\mathcal{E}_i^\delta[v_i]$ is given by

$$\begin{aligned} 0 &= \langle \delta_{\psi_i} \mathcal{E}_i^\delta, \theta \rangle \\ &= \frac{1}{n} \sum_{k=1}^n \int_{\mathcal{O}_k} W_{,A}(\mathcal{D}(\psi_i \circ \phi_k)) : (\mathcal{D}\theta \circ \phi_k \mathcal{D}\phi_k) \, dx - \delta \int_{\mathcal{S}} \sigma_i \nu[\mathcal{S}] \cdot \theta \, da \\ &= \int_{\mathcal{O}} \sigma[\delta v_i] : \mathcal{D}\theta \, dx - \delta \int_{\mathcal{S}} \sigma_i \nu[\mathcal{S}] \cdot \theta \, da \\ &= \int_{\mathcal{S}} ((\sigma[\delta v_i] - \delta\sigma_i)\nu[\mathcal{S}]) \cdot \theta \, da - \int_{\mathcal{O}} \operatorname{div} \sigma[\delta v_i] \cdot \theta \, dx \end{aligned}$$

for all test functions θ , where we denote by

$$n\sigma[\delta v_i] := \sum_{k=1}^n W_{,A}((I + \delta\mathcal{D}v_i)\mathcal{D}\phi_k \circ \phi_k^{-1}) \operatorname{cof} \mathcal{D}(\phi_k^{-1})$$

the first Piola–Kirchhoff stress tensor on the compound object \mathcal{O} . Hence, as optimality condition for v_i we indeed obtain

$$\operatorname{div} \sigma[\delta v_i] = 0 \quad \text{in } \mathcal{O}, \quad \sigma[\delta v_i]\nu[\mathcal{S}] = \delta\sigma_i\nu[\mathcal{S}] \quad \text{on } \mathcal{S}.$$

The stress tensor $\sigma[\delta v_i]$ here effectively reflects an average of all first Piola–Kirchhoff stresses in the n deformed configurations $\phi_i(\mathcal{O}_i)$ for $i = 1, \dots, n$. This can be seen by noting that

$$\begin{aligned} \sigma[\delta v_i] &= \frac{1}{n} \sum_{i=1}^n [W_{,A}(\mathcal{D}(\psi_i \circ \phi_k)) \operatorname{cof} \mathcal{D}\phi_k^{-1}] \circ (\phi_k)^{-1} \\ &= \left(\frac{1}{n} \sum_{i=1}^n [W_{,A}(\mathcal{D}(\psi_i \circ \phi_k)) \operatorname{cof} (\mathcal{D}(\psi_i \circ \phi_k))^{-1}] \circ (\psi_i \circ \phi_k)^{-1} \right) \circ \psi_i \operatorname{cof} \mathcal{D}\psi_i \\ &= \left(\frac{1}{n} \sum_{i=1}^n \sigma_k[\delta v_i] \right) \circ (\operatorname{id} + \delta v_i) \operatorname{cof} (\mathcal{D}(\operatorname{id} + \delta v_i)), \end{aligned}$$

where the $\sigma_k[\delta v_i]$ are indeed the Cauchy stresses of the different objects \mathcal{O}_k when deformed into $(\text{id} + \delta v_i)(\mathcal{O}) = ((\text{id} + \delta v_i) \circ \phi_k)(\mathcal{O}_k)$.

Since we are interested in the case of an infinitesimally small δ , we may expand the nonlinear energy about $\delta = 0$ and obtain, up to second order,

$$\begin{aligned}
 \mathcal{E}_i^\delta[v_i] &\doteq \frac{1}{n} \sum_{k=1}^n \left[\mathcal{W}[\mathcal{O}_k, \phi_k] + \delta \int_{\mathcal{O}_k} W_{,A}(\mathcal{D}\phi_k) : \mathcal{D}(v_i \circ \phi_k) \, dx \right. \\
 &\quad \left. + \frac{\delta^2}{2} \int_{\mathcal{O}_k} \langle W_{,AA}(\mathcal{D}\phi_k), \mathcal{D}(v_i \circ \phi_k), \mathcal{D}(v_i \circ \phi_k) \rangle \, dx \right] - \delta^2 \int_{\mathcal{O}} \sigma_i : \mathcal{D}v_i \, dx \\
 &= \frac{1}{n} \sum_{k=1}^n \mathcal{W}[\mathcal{O}_k, \phi_k] + \delta \int_{\mathcal{O}} \left(\frac{1}{n} \sum_{i=1}^n \sigma_k \right) : \mathcal{D}v_i \, dx \\
 &\quad + \frac{\delta^2}{2} \frac{1}{n} \sum_{i=1}^n \int_{\mathcal{O}_k} \langle W_{,AA}(\mathcal{D}\phi_k), \mathcal{D}(v_i \circ \phi_k), \mathcal{D}(v_i \circ \phi_k) \rangle \, dx - \delta^2 \int_{\mathcal{O}} \sigma_i : \mathcal{D}v_i \, dx \\
 &= \frac{1}{n} \sum_{k=1}^n \mathcal{W}[\mathcal{O}_k, \phi_k] + \delta \int_{\mathcal{S}} \left(\frac{1}{n} \sum_{i=1}^n \sigma_k \nu[\mathcal{S}] \right) \cdot v_i \, da - \delta \int_{\mathcal{O}} \left(\frac{1}{n} \sum_{i=1}^n \text{div} \sigma_k \right) \cdot v_i \, dx \\
 &\quad + \frac{\delta^2}{2} \frac{1}{n} \sum_{i=1}^n \int_{\mathcal{O}_k} \langle W_{,AA}(\mathcal{D}\phi_k), \mathcal{D}(v_i \circ \phi_k), \mathcal{D}(v_i \circ \phi_k) \rangle \, dx - \delta^2 \int_{\mathcal{O}} \sigma_i : \mathcal{D}v_i \, dx \\
 &= \frac{1}{n} \sum_{k=1}^n \mathcal{W}[\mathcal{O}_k, \phi_k] + \delta^2 \int_{\mathcal{O}} \frac{1}{2} \langle \mathbf{C}, \mathcal{D}v_i, \mathcal{D}v_i \rangle - \sigma_i : \mathcal{D}v_i \, dx
 \end{aligned}$$

for the in general inhomogeneous and anisotropic elasticity tensor

$$\mathbf{C} = \frac{1}{n} \sum_{k=1}^n \left(\frac{1}{\det \mathcal{D}\phi_k} \mathcal{D}\phi_k W_{,AA}(\mathcal{D}\phi_k) \mathcal{D}\phi_k^T \right) \circ \phi_k^{-1}.$$

This elasticity tensor takes into account the prestress of the compound configuration based on the combination of all deformations ϕ_k on the input objects \mathcal{O}_k for $k = 1, \dots, n$. Hence, we obtain v_i as the solution of the linear elasticity problem

$$\text{div}(\mathbf{C} \mathcal{D}v_i) = 0 \text{ in } \mathcal{O}, \quad \mathbf{C} \mathcal{D}v_i \nu[\mathcal{S}] = \sigma_i \nu[\mathcal{S}] \text{ on } \mathcal{S}$$

under the constraints $\int_{\mathcal{O}} v_i \, dx = 0$ and $\int_{\mathcal{O}} x \times v_i \, dx = 0$.

Note that for numerical implementation, instead of using the above form of the elasticity tensor, it is more suitable to implement the bilinear form

$$(u, v) \mapsto \frac{1}{n} \sum_{k=1}^n \int_{\mathcal{O}_k} \langle W_{,AA}(\mathcal{D}\phi_k), \mathcal{D}(u \circ \phi_k), \mathcal{D}(v \circ \phi_k) \rangle \, dx$$

directly without first applying the transformation rule.

The nonlinear formulation of the energy \mathcal{E}_i^δ is advantageous because it is of the same form as the averaging energy, but it is computationally cumbersome. For the computation of v_i , we will thus in general use the linearised elastic partial differential equations. However, if a strong amplification of v_i is required for visualisation of the shape variations, it might still be preferable to minimise the nonlinear energy if simple upscaling of the linear solution does not yield meaningful shape variations (due to the breakdown of the linear approximation).

5.1.2 Boundary stresses as alternative linearisations

The displacements v_i have the advantage that they may be interpreted as a direct variation of the average shape \mathcal{S} . However, instead of on these displacements, a covariance analysis for the shapes $\mathcal{S}_1, \dots, \mathcal{S}_n$ could as well be performed on the Cauchy boundary stresses $\sigma_i\nu[\mathcal{S}]$ themselves, which we have already identified as being characteristic of the corresponding input shapes \mathcal{S}_i (under the assumption that the relation between stresses and deformations is one-to-one, which is true at least locally, just as the exponential map on a Riemannian manifold is also only locally bijective). However, both choices are basically equivalent since the relationship between $\sigma_i\nu[\mathcal{S}]$ and v_i is linear and bijective so that switching between them only requires a change of the covariance metric (see next section).

Indeed, the partial differential equation and constraints for v_i are linear in v_i as well as $\sigma_i\nu[\mathcal{S}]$, and $\operatorname{div}(\mathbf{CD}v_i) = 0$ is basically the Laplace equation with Neumann boundary data $\mathbf{CD}v_i\nu[\mathcal{S}] = \sigma_i\nu[\mathcal{S}]$, which has a solution in $W^{1,2}(\mathcal{O})$ (unique up to a linearised rigid body motion, that is, an affine displacement with skew-symmetric matrix representation, which is fixed by the conditions of zero mean displacement and angular momentum) if the solvability condition

$$\int_{\mathcal{O}} \operatorname{div}(\mathbf{CD}v_i) \, dx = \int_{\mathcal{S}} \mathbf{CD}v_i\nu[\mathcal{S}] \, da$$

is fulfilled. This is indeed the case, since the left-hand side is zero, and due to the equilibrium of forces we must have $\int_{\mathcal{S}} \mathbf{CD}v_i\nu[\mathcal{S}] \, da = \int_{\mathcal{S}} \sigma_i\nu[\mathcal{S}] \, da = 0$.

That the solution v_i is unique in $W^{1,2}(\mathcal{O})$ up to a linearised rigid body motion can be easily seen as follows (assuming \mathcal{O} to be open and connected). If there were two solutions v, \tilde{v} , we would obtain $\operatorname{div}(\mathbf{CD}q) = 0$ in \mathcal{O} and $(\mathbf{CD}q)\nu[\mathcal{S}] = 0$ on \mathcal{S} for $q := v - \tilde{v} \in W^{1,2}(\mathcal{O})$. However, then the Dirichlet integral

$$\int_{\mathcal{O}} \operatorname{div}(q^T(\mathbf{CD}q)) \, dx$$

could be rewritten using either the divergence theorem or the product rule to obtain

$$\begin{aligned} 0 &= \int_{\mathcal{S}} q^T(\mathbf{C}\mathcal{D}q)\nu[\mathcal{S}] da = \int_{\mathcal{O}} \operatorname{div}(q^T(\mathbf{C}\mathcal{D}q)) dx \\ &= \int_{\mathcal{O}} \mathcal{D}q : (\mathbf{C}\mathcal{D}q) dx + \int_{\mathcal{O}} q^T \operatorname{div}(\mathbf{C}\mathcal{D}q) dx = \int_{\mathcal{O}} \mathcal{D}q : (\mathbf{C}\mathcal{D}q) dx. \end{aligned}$$

The elasticity tensor \mathbf{C} is positive semi-definite, and its eigenspace corresponding to the zero eigenvalue is the space of skew-symmetric matrices. We therefore obtain $\mathcal{D}q : (\mathbf{C}\mathcal{D}q) = 0$ and thus (in the case $d = 3$)

$$\mathcal{D}q = \begin{pmatrix} 0 & a(x_1, x_2, x_3) & b(x_1, x_2, x_3) \\ -a(x_1, x_2, x_3) & 0 & c(x_1, x_2, x_3) \\ -b(x_1, x_2, x_3) & -c(x_1, x_2, x_3) & 0 \end{pmatrix}$$

almost everywhere. Using Schwarz's theorem in a distributional sense, we obtain $0 = \frac{\partial}{\partial x_2} \frac{\partial q_1}{\partial x_1} = \frac{\partial}{\partial x_1} \frac{\partial q_1}{\partial x_2} = \frac{\partial a}{\partial x_1}$ almost everywhere and the analogue for the other derivatives so that finally $a = a(x_3)$, $b = b(x_2)$, $c = c(x_1)$ almost everywhere by the fundamental lemma of variational calculus. This implies $q_1 = \kappa_1 + \kappa_2 x_2 x_3 + \kappa_3 x_2 + \kappa_4 x_3$ for some constants $\kappa_1, \dots, \kappa_4$ and similar results for q_2 and q_3 , and by the condition of a skew-symmetric gradient we obtain $q = q_0 + A(x_1, x_2, x_3)^T$ for a constant vector q_0 and a skew-symmetric matrix A .

5.2 Covariance analysis

In order to perform a principal component analysis on a set of shapes $\mathcal{S}_1, \dots, \mathcal{S}_n$, in the previous section we sought for representatives of the shapes in a linear vector space and chose the displacements $v_i : \mathcal{O} \rightarrow \mathbb{R}^d$, $i = 1, \dots, n$. These displacements v_i reflect the variations of the average shape induced by a modulation of the boundary stresses $\sigma_i \nu[\mathcal{S}]$ from the deformations ϕ_i of the input shapes \mathcal{S}_i into the average shape \mathcal{S} . Owing to the linear relationship between stresses $\sigma_i \nu[\mathcal{S}]$ and displacements v_i , the pointwise stress balance, $0 = \sum_{i=1}^n \sigma_i \nu[\mathcal{S}]$ (compare Figure 4.5), implies the v_i to be centred,

$$0 = \sum_{i=1}^n v_i,$$

so that a principal component analysis on these displacements can directly be applied after the definition of a suitable inner product $g(v, \tilde{v})$ for displacements $v, \tilde{v} : \mathcal{O} \rightarrow \mathbb{R}^d$.

Note that the metric $g(\cdot, \cdot)$ induces a metric $\tilde{g}(\sigma \nu[\mathcal{S}], \tilde{\sigma} \nu[\mathcal{S}]) := g(v, \tilde{v})$ on the associated boundary stresses. Hence, the covariance analysis presented in the following can also be considered as a corresponding analysis directly on boundary stresses $\sigma_1 \nu[\mathcal{S}], \dots, \sigma_n \nu[\mathcal{S}]$. Indeed, the symmetry and bilinearity of $\tilde{g}(\cdot, \cdot)$ follow directly from

the symmetry and bilinearity of $g(\cdot, \cdot)$ as well as the linearity of the relation between stresses and displacements, and the positive definiteness of $\tilde{g}(\cdot, \cdot)$ follows from the positive definiteness of $g(\cdot, \cdot)$ and the injectivity of the map $\sigma\nu[\mathcal{S}] \mapsto v$, which assigns each boundary stress $\sigma\nu[\mathcal{S}]$ a corresponding displacement v with zero mean and angular momentum.

We will consider two different inner products g on the displacements:

- *The L^2 -product.* Given two square-integrable displacements v, \tilde{v} , we define

$$g(v, \tilde{v}) := \int_{\mathcal{O}} v \cdot \tilde{v} \, dx.$$

This product weights local displacements equally on the whole compound object \mathcal{O} . It pronounces shape variations with large displacements even though they are energetically cheap (for example, a rotation of some structure around a joint).

- *The Hessian of the energy as inner product.* Different from the L^2 -metric, we now measure displacement gradients in a non-homogeneous way. Precisely, we define

$$g(v, \tilde{v}) := \int_{\mathcal{O}} \mathbf{C}\mathcal{D}v : \mathcal{D}\tilde{v} \, dx$$

for displacements v, \tilde{v} with square-integrable gradients. Hence, the contribution to the inner product is larger in areas of the compound object which are in a significantly stressed configuration.

The chosen inner product $g(\cdot, \cdot)$ induces a covariance operator

$$\mathbf{Cov} \, v := \frac{1}{n} \sum_{k=1}^n g(v, v_k) v_k$$

on the space of displacements. It is obviously symmetric positive definite with respect to $g(\cdot, \cdot)$ on $\text{span}(v_1, \dots, v_n)$ and can be diagonalised, yielding eigendisplacements w_k and corresponding eigenvalues λ_k , $k = 1, \dots, n$, with

$$\mathbf{Cov} \, w_k = \lambda_k w_k.$$

These can be obtained by diagonalising the symmetric matrix $\frac{1}{n} (g(v_i, v_j))_{i,j=1,\dots,n} = \mathbf{O}\mathbf{\Lambda}\mathbf{O}^T$, where $\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_n)$ and \mathbf{O} is orthogonal. The eigendisplacements are then recovered as $w_k = \frac{1}{\sqrt{\lambda_k}} \sum_{j=1}^n O_{jk} v_j$. They can be regarded as the principal modes of the variation of \mathcal{O} and thus of $\mathcal{S} = \partial\mathcal{O}$, given shapes $\mathcal{S}_1, \dots, \mathcal{S}_n$, where the eigenvalues indicate the actual strength of the variations. The resulting modes of variation can easily be visualised via a scalar modulation δw_k for varying values of δ , which is associated with a corresponding modulation of the underlying stresses $\delta n \mathbf{C}\mathcal{D}w_k \nu[\mathcal{S}]$. If an amplified



Figure 5.1: The two dominant modes (right) for four different shapes (left) demonstrate that the principal component analysis properly captures strong geometric nonlinearities.

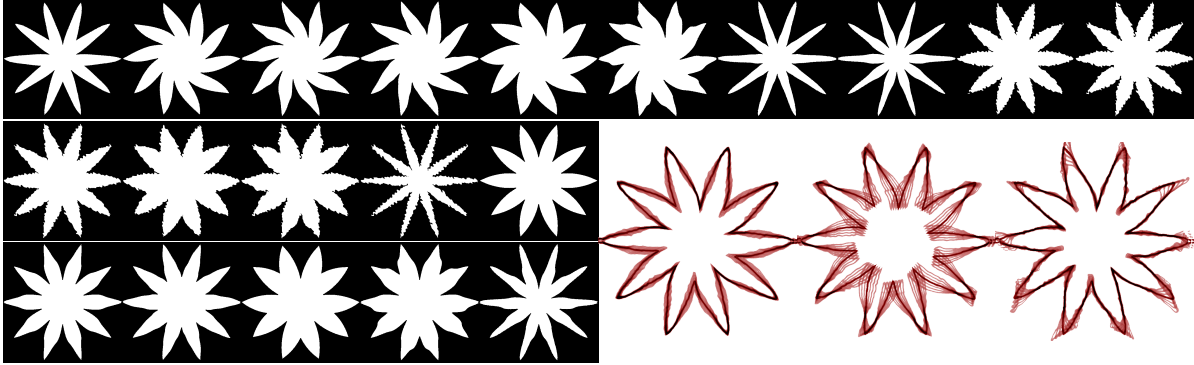


Figure 5.2: Original shapes from Figure 4.10 and their first three modes of variation with ratios $\frac{\lambda_i}{\lambda_1}$ of 1, 0.20, and 0.05.

visualisation of the modes is required, it is again preferable to depict displacements w_δ^k which are defined as minimisers of the nonlinear variational energy

$$\frac{1}{n} \sum_{i=1}^n \mathcal{W}[\mathcal{O}_i, (\text{id} + w_\delta^k) \circ \phi_i] - \delta^2 \int_{\mathcal{S}} \mathbf{CD}w_k \nu[\mathcal{S}] \cdot w_\delta^k \, da.$$

This covariance analysis properly takes into account the usually strong geometric non-linearity in shape analysis via the transfer of geometric shape variation to elastic stresses on the average shape, based on paradigms from nonlinear elasticity. This is illustrated in Figure 5.1 for the L^2 -metric as underlying inner product. Similarly, in Figure 5.2, a larger set of 20 binary images “device7” from the MPEG7 shape database serves as input shapes. Apparently, the first principal component is given by a thickening or thinning of the lobes, accompanied by a change of indentation depth between them. The second mode obviously corresponds to bending the lobes, and the third mode represents local changes at the tips: A sharpening and orientation of neighbouring tips towards each other, originating, for example, from the sixth or the second last input shape.

Displacements (or stresses) are interpreted as the proper linearisation of shapes. In abstract terms, either the space of displacements or stresses can be considered as the tangent space to shape space at the average shape, where the identification of displacements and stresses via a linearised elasticity problem provides a suitable physical interpretation of stresses as modes of shape variation.

5.2.1 Impact of the covariance metric

Naturally, the modes of variation depend on the chosen inner product for the displacements v_i , for which there are various possibilities. In order to be physically meaningful, we should here employ displacements v_i that are obtained by applying the scaled boundary stresses $\delta n\sigma_i\nu[\mathcal{S}]$ to the compound object \mathcal{O} which is composed of all deformed input objects $\phi_i(\mathcal{O}_i)$. If instead we apply the stresses $\delta n\sigma_i\nu[\mathcal{S}]$ to an object which just looks like the average shape, but does not contain the information how strongly the input shapes had to be deformed to arrive at the average, then we obtain different displacements v_i and thus a different result as shown in Figure 5.3, where we compare different inner products for a set of six locally bent versions of an originally straight bar.

The first computation in Figure 5.3 shows the dominant modes of shape variation for the L^2 -metric, if the additional stresses $\delta n\sigma_i\nu[\mathcal{S}]$ are assumed to act on an unstressed material in order to obtain the displacements v_i (as opposed to interpreting the average object \mathcal{O} as a compound object). The second computation employs the L^2 -metric on the displacements v_i , which this time are obtained on the basis of interpreting \mathcal{O} as the compound object of all deformed input objects. Finally, the last computation has been performed using the Hessian of the nonlinear elastic energy as the inner product.

The different behaviour of the principal component analysis, when \mathcal{O} is not regarded as a compound object, can be attributed to the following fact. Regions which were more heavily deformed than others need higher stresses to be deformed even further. Therefore, these regions exhibit much stronger variations if the inner product is based on a non-compound object (which is not already prestressed) than on a compound, prestressed object.

Concerning the difference between the L^2 -product and the Hessian-based inner product, let us note that in contrast to the L^2 -metric, the metric defined via the Hessian of the elastic energy captures differences in the deformation in exactly those regions where the deformation takes place. Furthermore, a clearer separation of mechanically separated regions is observed compared to the L^2 -metric.

5.2.2 Impact of the nonlinear elastic constitutive law

The particular choice of the nonlinear elastic energy density also has a considerable effect on the average shape and its modes of variation. Figure 5.4 has been obtained using the hyperelastic energy density $\hat{W}(I_1, I_2, I_3) = \frac{\mu}{2}I_1^2 + \frac{\lambda}{4}I_3^2 - (\mu + \frac{\lambda}{2})\log I_3 - \mu - \frac{\lambda}{4}$, where μ and λ are the factors of length and volume change penalisation for small deformations, respectively. A low penalisation of volume changes apparently leads to local compression or inflation at the dumbbell ends (left), while for deformations with a strong volume change penalisation (right), material is squeezed from one end to the other, which becomes especially apparent in the second and third mode of variation. Note that in the first mode, the volume of one dumbbell end shrinks while the volume of the other end increases, whereas for the second mode both ends change equally. This

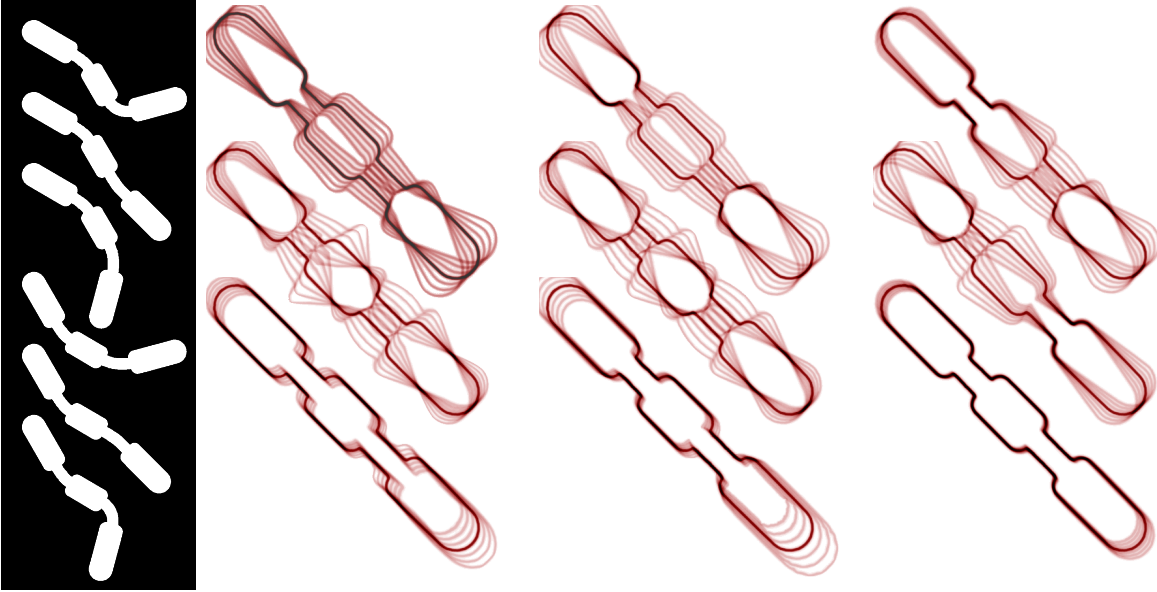


Figure 5.3: The first three dominant modes of variation for the six input shapes on the left obviously depend on the employed metric: The left column depicts the modes belonging to the L^2 -metric on displacements of a non-prestressed object (with ratios $\frac{\lambda_i}{\lambda_1}$ of 1, 0.23, and 0.07), the middle column corresponds to the L^2 -metric on displacements of the proper compound object (with ratios $\frac{\lambda_i}{\lambda_1}$ of 1, 0.28, and 0.03), the right column represents the results for the energy Hessian based metric on displacements of the compound object (with ratios $\frac{\lambda_i}{\lambda_1}$ of 1, 0.61, and 0.24).

is a very illustrative example of the orthogonality of the decomposition of the tangent space into subspaces according to the different modes of variation. Here, the underlying metric is the one based on the Hessian of the energy.

5.2.3 Elastic versus Riemannian shape analysis

As has already been mentioned earlier, there are fundamental differences between an elasticity-based structure on shape space and a Riemannian one.

A Riemannian structure is induced by tangent spaces with an inner product at each shape $\mathcal{S} = \partial\mathcal{O}$ or object \mathcal{O} , where the inner product $g(v, \tilde{v})$ between two displacement fields $v, \tilde{v} : \mathcal{O} \rightarrow \mathbb{R}^d$ in the same tangent space is typically defined via an elliptic operator such as

$$g(v, \tilde{v}) = \int_{\mathcal{O}} \frac{\lambda}{2} \operatorname{tr} \epsilon[v] \operatorname{tr} \epsilon[\tilde{v}] + \mu \epsilon[v] : \epsilon[\tilde{v}] \, dx,$$

where $\epsilon[v] = \frac{1}{2}(\mathcal{D}v^T + \mathcal{D}v)$, $\epsilon[\tilde{v}] = \frac{1}{2}(\mathcal{D}\tilde{v}^T + \mathcal{D}\tilde{v})$. The distance between any two shapes

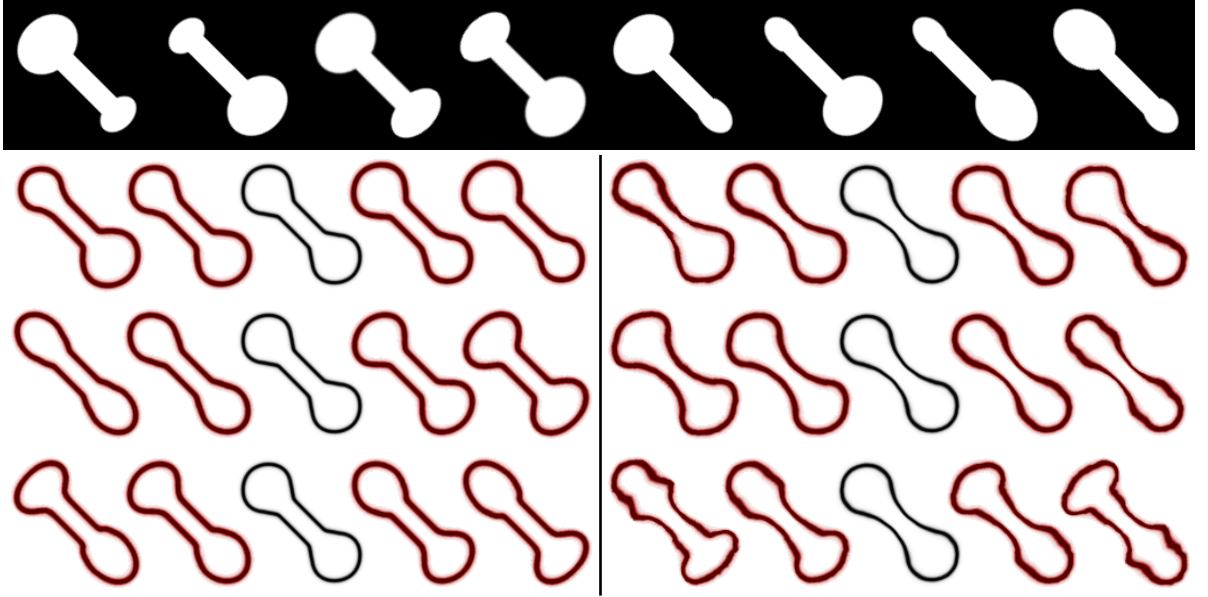


Figure 5.4: The first three modes of variation for eight dumbbell shapes, left for a 100 times stronger penalization of length changes than of volume changes (with ratios $\frac{\lambda_i}{\lambda_1}$ of 1, 0.22, and 0.05), right for the reverse (with ratios $\frac{\lambda_i}{\lambda_1}$ of 1, 0.41, and 0.07).

$\mathcal{S}_1, \mathcal{S}_2$ (or corresponding objects) is then defined as the minimum geodesic path length,

$$\min_{v_t} \int_0^1 \sqrt{g(v_t, v_t)} dt,$$

where $v_t(t)$ for $t \in [0, 1]$ is an element of the tangent space to the shape space at $\mathcal{O}(t)$ such that the path $\mathcal{O}(t)$ is generated by v_t and connects \mathcal{O}_1 and \mathcal{O}_2 .

In the elastic setting, as an analogue to a tangent space at a shape $\mathcal{S} = \partial\mathcal{O}$ we employed the space of all boundary stresses on \mathcal{S} or displacements on \mathcal{O} , and we also imposed an inner product on this tangent space. There are fundamental differences, though. Elements of our “tangent” space do not generate any path in shape space; due to the fundamental axiom of elasticity, there is not even a physically meaningful notion of a path in shape space: The state and energy of an elastically deformed object are independent of how they were attained. This issue is also connected to the nonlocal interpretation of a “tangent” vector in the elastic approach. While a tangent vector on a Riemannian manifold describes the local rate of change of a shape along a path, our tangent vector is not associated with the local direction of a path but instead with one particular (possibly very distant) shape in shape space. Furthermore, the inner product on the space of boundary stresses or displacements in general depends on all shapes that are considered during the statistical analysis.

On a Riemannian manifold, the exponential map allows to describe geodesics from

an averaged shape \mathcal{S} to the input shapes \mathcal{S}_k via $\mathcal{S}_k = \exp_{\mathcal{S}}(\tilde{v}_k)$ for some tangent vector \tilde{v}_k to shape space at \mathcal{S} . Hence, a covariance analysis will be performed on the tangent vectors $\tilde{v}_1, \dots, \tilde{v}_n$ with respect to the Riemannian metric $g(\cdot, \cdot)$. In the strictly elastic setup, instead, the stresses $\sigma_k \nu[\mathcal{S}]$ or their induced displacements v_k play the role of the \tilde{v}_k , representing the impact of \mathcal{S}_k on the average shape \mathcal{S} .

The lack of paths in the elasticity-based shape space, whose length can be measured, has the effect that the shape space is in general not metrisable. We have already emphasised several times that the elastic deformation energy of a deformation between two shapes (or its square root) is neither symmetric nor does it satisfy the triangle inequality. Both properties only hold in the limit of infinitesimal deformations, the regime of linearised elasticity, where the elastic energy is actually quadratic. However, this fact provides a hint to the connection between the elastic and the Riemannian perspective. If we denote by $d_{\mathcal{W}}(\mathcal{S}_1, \mathcal{S}_2) = \inf_{\phi(\mathcal{O}_1)=\mathcal{O}_2} \sqrt{\mathcal{W}[\mathcal{O}_1, \phi]}$ the distance measure in the elasticity-based shape space, then we can obtain a metric by defining

$$d(\mathcal{S}_1, \mathcal{S}_2) = \inf_{\substack{n \in \mathbb{N} \\ \tilde{\mathcal{S}}_1, \dots, \tilde{\mathcal{S}}_n \\ \tilde{\mathcal{S}}_1 = \mathcal{S}_1, \tilde{\mathcal{S}}_n = \mathcal{S}_2}} \sum_{i=1}^{n-1} d_{\mathcal{W}}(\tilde{\mathcal{S}}_i, \tilde{\mathcal{S}}_{i+1}),$$

and this metric will be the Riemannian geodesic distance induced by viscous dissipation (see Chapter 6). This relation will form the basic idea of a variational time discretisation of geodesic paths in the next chapter. As the above definition of the metric indicates, the crucial difference between the elastic and the viscous approach is that in the Riemannian setting, shapes are at each time in an unstressed state, while in the elastic setting, the deformed shapes bear stresses.

The above-mentioned conceptual differences are reflected in a different behaviour. If we regard shapes from a flow-oriented perspective, then a viscous approach would be more appropriate. However, the elastic approach is favourable for rather rigid, more stable shapes, since it prevents locally strong isometry violation. An example is provided in Figure 5.5: The input shapes are regarded as two versions of an object that may have none, one, or two pins at more or less stable positions. Both pins are apparently not interpreted as shifted versions of each other since a shifting deformation would cost too much energy. However, if the material was visco-plastic, a horizontal shift of each pin would be easier and result in an average shape with just one centred pin and its variation being a sideward movement. This corresponds to a completely different perception of the input shapes.

The strong local rigidity and isometry preservation of the elasticity concept becomes particularly evident in Figures 5.1 and 5.6, where non-isometric deformations are concentrated only at joints. This holds true already in the case of an underlying L^2 -metric as inner product as it is taken into account here.



Figure 5.5: Average and variation (right) for two shapes with pins at different positions (left). The pins are not interpreted as shifted versions of each other.

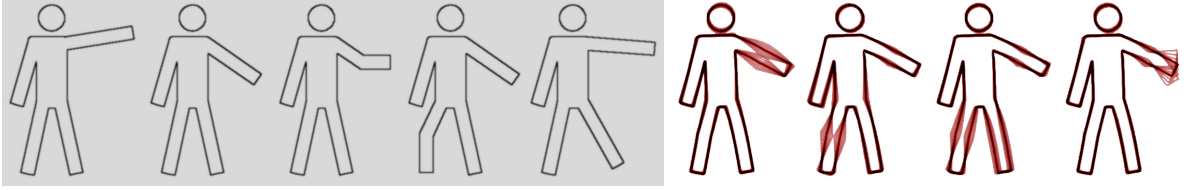


Figure 5.6: A set of input shapes (compare Figure 4.1) and their modes of variation with ratios $\frac{\lambda_i}{\lambda_1}$ of 1, 0.22, 0.15, and 0.06.

5.3 Numerical implementation

As in the previous chapter, we replace the given objects \mathcal{O}_i by an elastic material which covers the whole computational domain $\Omega = [0, 1]^d$, but whose stiffness is reduced by the factor 10^{-4} on $\Omega \setminus \mathcal{O}_i$. Here, again, the characteristic function of an object \mathcal{O} is approximated by a smoothed version $\chi_{\mathcal{O}}^\varepsilon$ with interface thickness ε so that the linear representation v_i of shape \mathcal{S}_i is obtained as the minimiser of the functional

$$\begin{aligned} \mathcal{E}_i^{\delta, \varepsilon}[v_i] &= \frac{1}{n} \sum_{k=1}^n \mathcal{W}^\varepsilon[\mathcal{O}_k, (\text{id} + \delta v_i) \circ \phi_k] - \delta^2 \int_{\Omega} ((1 - \delta)\chi_{\mathcal{O}}^\varepsilon + \delta)\sigma_i : \mathcal{D}v_i \, dx \\ &= \frac{1}{n} \sum_{k=1}^n \int_{\Omega} ((1 - \delta)\chi_{\mathcal{O}}^\varepsilon + \delta) [W(\mathcal{D}((\text{id} + \delta v_i) \circ \phi_k)) - \delta^2 \sigma_i : \mathcal{D}v_i] \, dx \end{aligned}$$

or (for an infinitesimal δ) its quadratic part

$$\delta^2 \int_{\Omega} ((1 - \delta)\chi_{\mathcal{O}}^\varepsilon + \delta) \left[\frac{1}{2} \langle \mathbf{C}, \mathcal{D}v_i, \mathcal{D}v_i \rangle - \sigma_i : \mathcal{D}v_i \right] \, dx.$$

We employ the same spatial discretisation as in the previous chapter with continuous, multilinear finite elements on a regular grid and Gaussian quadrature of third order on each grid cell. We denote by V_i the finite element approximation to v_i , and the vector of nodal values by $\mathbf{V}_i \in \mathbb{R}^{dN}$, where N denotes the number of grid vertices. In order to obtain V_i as the finite element function that minimises $\mathcal{E}_i^{\delta, \varepsilon}[V_i]$, we perform a gradient descent with Armijo line search just as explained in detail in the previous chapter. If the displacements V_i are to be obtained as the solution to the linear elasticity problem,

then we simply use a conjugate gradient iteration to solve the system of linear equations

$$L_{\mathbf{C}} \mathbf{V}_i = b_i,$$

where the stiffness operator encoding the elastic tensor \mathbf{C} is given by

$$L_{\mathbf{C}} = \left(\frac{1}{n} \sum_{k=1}^n \int_{\Omega} ((1-\delta)\chi_{\mathcal{O}_k}^\varepsilon + \delta) \langle W_{,AA}(\mathcal{D}\phi_k), \mathcal{D}((\varphi_i e_l) \circ \phi_k), \mathcal{D}((\varphi_j e_m) \circ \phi_k) \rangle dx \right)_{\substack{(i,l),(j,m) \\ \in I_h \times \{1,\dots,d\}}}$$

for the canonical finite element basis functions φ_j , $j \in I_h$ (I_h being the index set of all grid nodes), and Euclidean basis vectors e_1, \dots, e_d , and where the right-hand side encodes the Neumann boundary conditions,

$$b_i = \left(\int_{\Omega} ((1-\delta)\chi_{\mathcal{O}}^\varepsilon + \delta) \sigma_i : \mathcal{D}(\varphi_j e_m) dx \right)_{(j,m) \in I_h \times \{1,\dots,d\}}.$$

The stiffness operator $L_{\mathbf{C}}$ is assembled by first computing $\phi_k(x)$ at each quadrature point x . Then, all those basis functions θ_i are identified whose support contains $\phi_k(x)$. Points which are displaced outside the computational domain are projected back onto its boundary. The corresponding evaluation of $\theta_i \circ \phi_k$ at x contributes to the assembly of the stiffness operator.

5.4 Validation and applications

In the following, the shape analysis approach is applied to collections of 2D and 3D shapes.

5.4.1 PCA for 2D input shapes

Some results of shape averages and corresponding dominant modes of shape variation for shapes of 2D objects are already depicted in Figures 5.1 to 5.6 as first illustrative examples. Especially Figures 5.1 and 5.6 show that—due to the invariance properties of the energy—isometries are locally preserved in the dominant modes of shape variation. In both examples, the average shape is represented by the dark line, whereas the light red lines signify deformations of the shape along the principal components. In Figure 5.6, we see the bending of the arm and the leg basically decoupled as the first two dominant modes of variation. The silhouette variations of raising the arm or the leg can only be obtained as linear combinations of the first and fourth or of the second and third mode of variation, respectively. This coupling is not too surprising, noting that the average has a slightly bent leg and arm so that the influence of all input shapes on the average also incorporates a straightening or bending of these limbs.

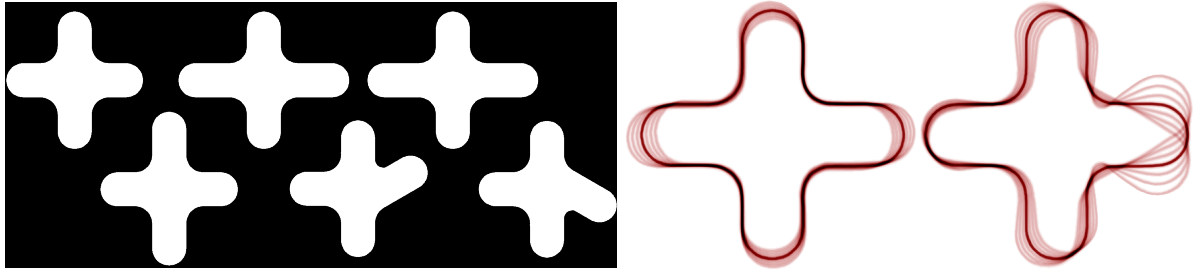


Figure 5.7: Six input shapes and their first two modes of variation with ratios $\frac{\lambda_i}{\lambda_1}$ of 1 and 0.34.

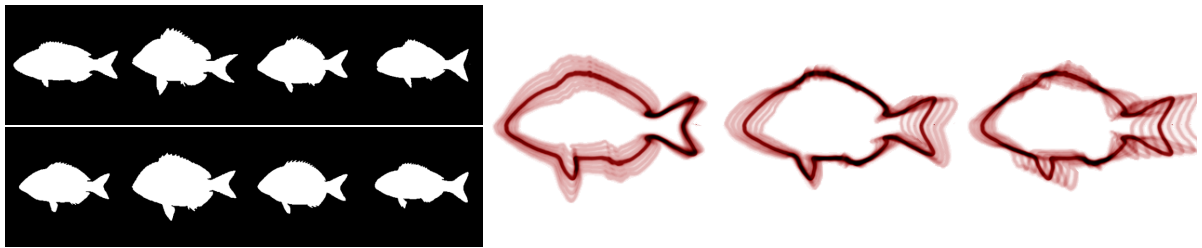


Figure 5.8: First three modes of variation with ratios $\frac{\lambda_i}{\lambda_1}$ of 1, 0.49, and 0.26 for the fish silhouettes from Figure 4.12.

In the following examples the covariance analysis is performed based on the metric induced by the Hessian of the elastic energy. The decoupling of shape variations becomes even more obvious in Figure 5.7, where we have more input shapes but fewer variation among them. The first mode describes the shortening of the horizontal and stretching of the vertical axis (or vice versa), whereas the second mode corresponds to bending the right branch of the cross-shape. The additional tilt of the deformed shapes in the second mode is due to the condition on the shape variations of zero angular momentum. The complete decoupling of bending and stretching is here achieved by including the cross with bent branch as well as its symmetric counterpart as input shapes. The average shape then only has straight branches so that the bending is invisible to the stretching modes.

The results from Figure 5.8 are shown for comparison with [31]. Interestingly, the obtained modes of variation differ slightly: In both cases, the first mode of variation is some kind of height variation of the mean fish (though locally, the variation looks different). While our second mode of variation is more or less an overall variation in fish length, especially pronounced at the tail fin, they obtain a combination of different local variations like tail fin thickness, pectoral fin length, and chest shape. Such a type of eigenmode in our computation only occurs as the third mode of variation.

A statistical analysis of the hand shapes in Figure 5.9 has also been performed in [43] and [59], where the shapes are represented as vectors of landmark positions. The average

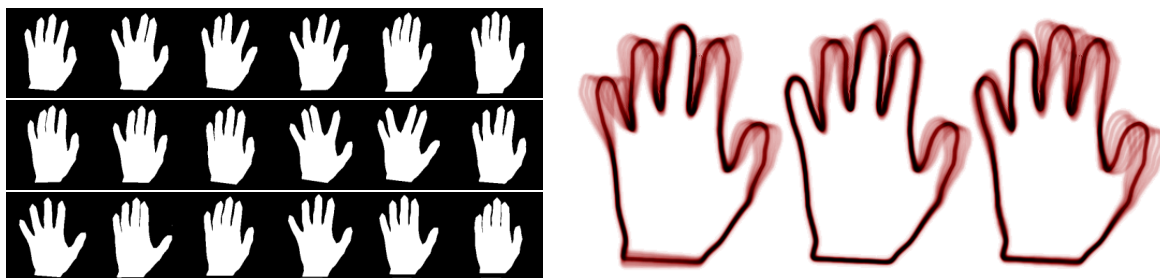


Figure 5.9: First three modes of variation with ratios $\frac{\lambda_i}{\lambda_1}$ of 1, 0.88, and 0.42 for 18 hand silhouettes from Figure 4.12.

and the modes of variation are quite similar, representing different ways of spreading the fingers.

5.4.2 PCA for image morphologies

Figure 5.10, using thorax CT scans as input images, shows that the approach also works for image morphologies instead of shapes. As in the previous chapter, the edge set of the images is considered as the corresponding shape. Hence, these shapes are usually significantly more complex and characterised by nested shape contours. In our example, the first mode of variation represents a variation of the chest size, the next mode corresponds to a change of heart and scapula shape, while the third mode mostly concerns the rib position. As for Figure 4.16, the input images were not segmented in advance, but simultaneously to the averaging procedure to exploit the stronger robustness of joint segmentation and registration. In this way, artifacts have been avoided that would appear otherwise due to the visibility of liver contours in only some of the input images. Note that the local shape variation at the sides of the chest in the second and third mode of variation originates from the visibility of the scapula in some input images.

5.4.3 PCA for 3D shapes

Next, let us investigate the dominant modes of variation of shapes in \mathbb{R}^3 , where the computation is based on the L^2 -metric. In our first 3D example we compute the first four modes of variation for the set of 48 kidney shapes from Figure 4.7 which are once more depicted in Figure 5.11. For all modes we show the average in the middle and its configurations after deformation according to the principal components. Local structures seem to be quite well represented and preserved during the averaging process and the subsequent covariance analysis compared to, for example, the PCA on kidney shapes in [57] where a medial representation is used. The second example takes the 24 foot-shapes from Figure 4.14 as input. It is doubtlessly difficult to analyse the shape variation solely on the basis of the colour-coding in Figure 4.14: We see modest variation at the toes and the heel as well as on the instep, but any correlation between these variations is difficult

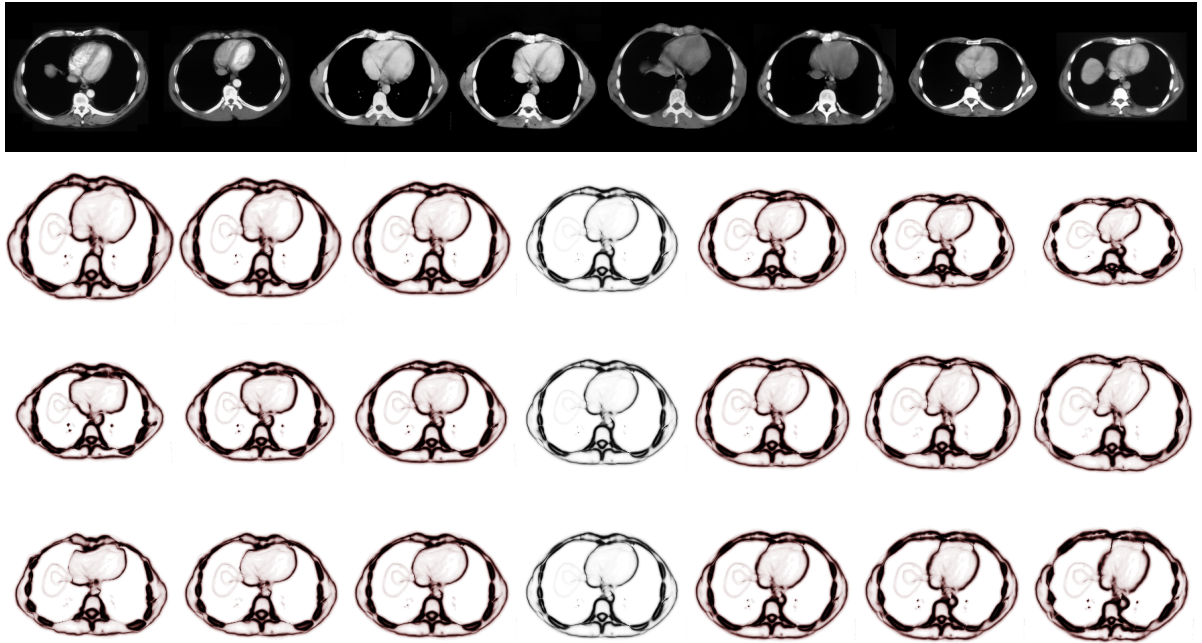


Figure 5.10: 8 thorax CT scans from different patients (top row) and their first three modes of variation with ratios $\frac{\lambda_i}{\lambda_1}$ of 1, 0.12, and 0.07. In each row, the middle shape corresponds to the average, and the shapes to its left and right visualise its variation according to a single principal mode. Note that the thin lines which can be seen left of the heart correspond to contours of the liver, which are only visible in the first and last input image.

to determine. The corresponding modes of variation in Figure 5.12, however, are quite intuitive. The first mode apparently represents changing foot lengths, the second and third mode belong to different variants of combined width and length variation, and the fourth to sixth mode correspond to variations in relative heel position, ankle thickness, and instep height.

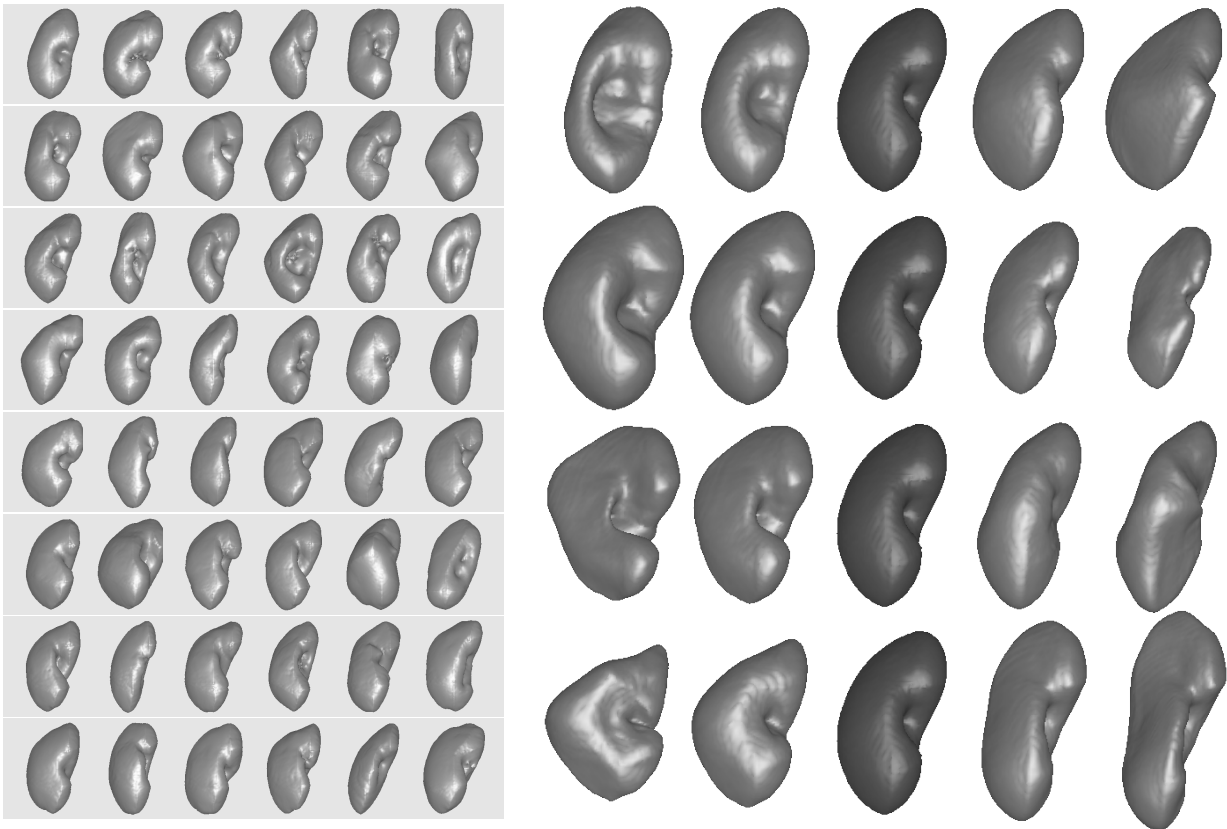


Figure 5.11: 48 input kidneys and their first four modes of variation with ratios $\frac{\lambda_i}{\lambda_1}$ of 1, 0.72, 0.37, and 0.31.

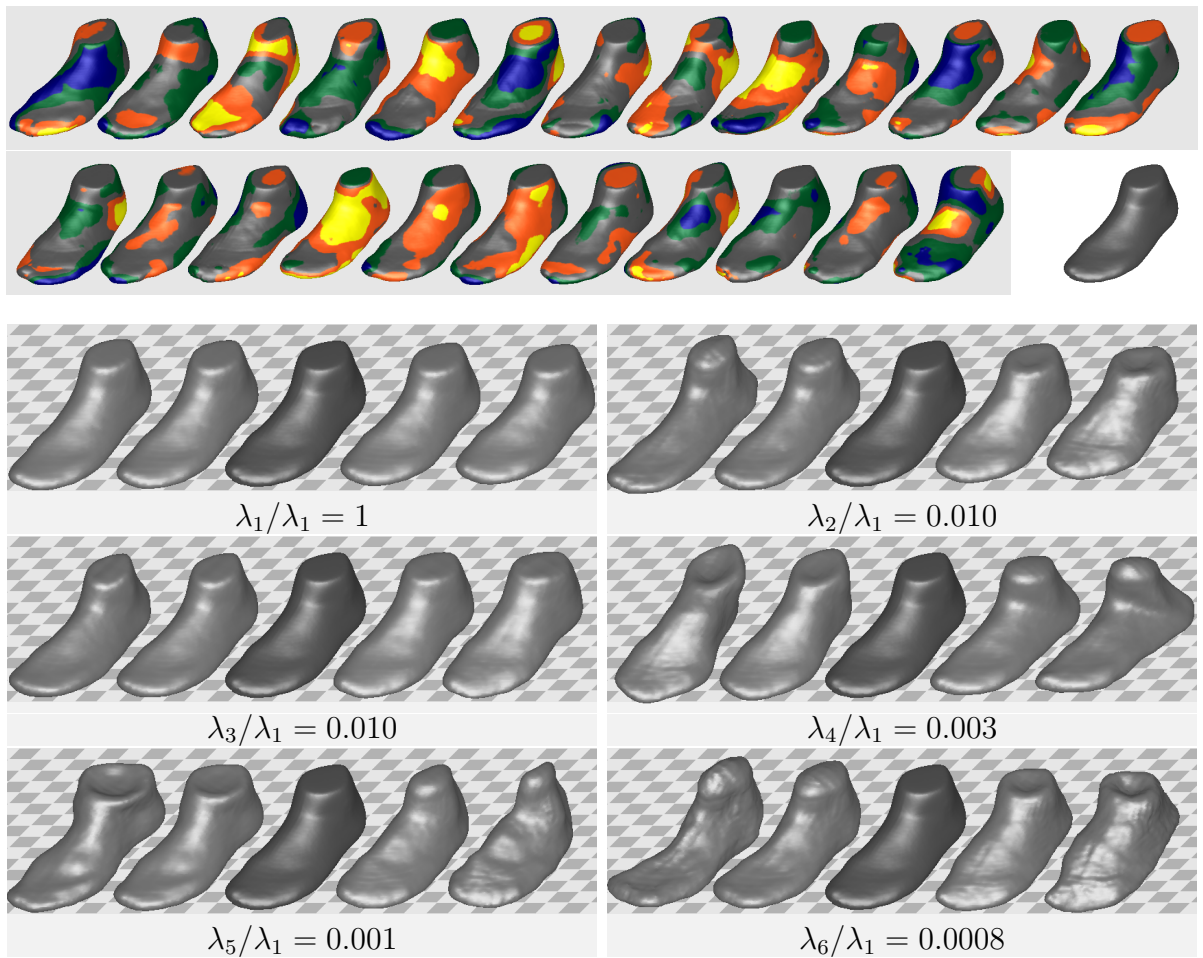


Figure 5.12: The 24 foot shapes from Figure 4.14 and their average as well as the first six dominant modes of variation.

6 Geodesics based on viscous flow and their variational time discretisation

In this chapter we will equip the space of sufficiently regular, deformable objects with a Riemannian structure, where we employ the viscous dissipation of a deformation as the Riemannian metric. We will compute geodesic paths between shapes which are represented by level set functions and possibly are of different topology. The computation will be based on a variational time discretisation which establishes a link between pairwise elastic shape matching and the Riemannian flow perspective on paths in shape space.

The concept of shortest paths in shape space is of interest especially in computer vision for the morphing of different shapes into each other (for example, two shapes that represent different configurations of the same object). The employed metric determines the properties of the resulting path and can be chosen to yield the most natural warping for the considered context. The lengths of the obtained geodesic paths can also be used as a distance measure for shape clustering.

Concerning the shape space modelling and the approximation of geodesics we will proceed as follows. We regard a path $\mathcal{S}(t) = \partial\mathcal{O}(t)$, $t \in [0, 1]$, in shape space as a time-continuous deformation of $\mathcal{O}(0)$ into $\mathcal{O}(1)$ such that $\mathcal{O}(t)$ is generated by a motion field $v(t) : \mathcal{O}(t) \rightarrow \mathbb{R}^d$, which represents the time derivative or the velocity field of the deformation. $\mathcal{O}(t)$ is considered to be made of a viscous material so that internal friction will occur during the deformation which results in viscous dissipation, that is, conversion of mechanical energy into heat. This viscous dissipation depends on the motion field $v(t)$, and it induces a Riemannian metric on the space of all motion fields on an object $\mathcal{O}(t)$ (which is the tangent space to shape space at $\mathcal{O}(t)$).

For the computation of geodesics, a straightforward discretisation of the corresponding motion field will neither yield rigid body motion invariance of the discrete geodesic nor a one-to-one correspondence between different shapes along the path so that we choose to discretise a geodesic as a sequence of specific pairwise matching problems for consecutive shapes on the time-discrete path. The rigid body motion invariance and one-to-one correspondence property are then inherited from the pairwise matching problems. This will allow to approximate geodesics with only few intermediate shapes.

We will first introduce the underlying physical modelling idea of paths in shape space in Section 6.1 and then introduce a variational time discretisation for the computation

of geodesics in Section 6.2. Next, the description of the shapes via level sets will be described in Section 6.3 before the numerical implementation will be specified in Section 6.4. Finally, we will show some applications in Section 6.5 and conclude in Section 6.6.

6.1 Paths in shape space generated by viscous deformation

As elsewhere in this thesis, we regard shapes \mathcal{S} as boundaries $\partial\mathcal{O}$ of deformable, sufficiently regular open object domains $\mathcal{O} \subset \mathbb{R}^d$. A path $\mathcal{S}(t)$, $t \in [0, 1]$, in shape space then corresponds to a path of objects $\mathcal{O}(t)$ with $\mathcal{S}(t) = \partial\mathcal{O}(t)$. We assume such a path to be generated by a sufficiently regular flow $\psi : [0, 1] \times \mathcal{O}(0) \rightarrow \mathbb{R}^d$ such that $\mathcal{O}(t) = \psi(t, \mathcal{O}(0))$, that is, $\psi(t, \cdot) : \mathcal{O}(0) \rightarrow \mathbb{R}^d$ is the deformation connecting $\mathcal{O}(0)$ with $\mathcal{O}(t)$. Instead of the flow ψ , one might equivalently specify the corresponding motion or velocity field $v(t) : \mathcal{O}(t) \rightarrow \mathbb{R}^d$ for each time $t \in [0, 1]$, which vanishes outside the shapes and integration of which returns the flow. The relation between flow and velocity field is obviously given by

$$v(t) = \frac{\partial\psi(t, \cdot)}{\partial t} \circ \psi(t, \cdot)^{-1}.$$

Via the above identification of paths in shape space with flows $\psi : [0, 1] \times \mathcal{O}(0) \rightarrow \mathbb{R}^d$, a motion field $v : \mathcal{O} \rightarrow \mathbb{R}^d$ can be regarded as a tangent vector to shape space at $\mathcal{S} = \partial\mathcal{O}$. We thus obtain a Riemannian shape space after defining a first fundamental form for all possible such motion fields v , and we choose this first fundamental form to be motivated by physics; in particular, we choose it to coincide with the viscous dissipation of the motion field v , that is, the rate at which mechanical energy is converted into heat due to internal friction. In different words, we pretend the object $\mathcal{O}(t)$ to be made of a real physical, viscous material so that friction occurs inside the material during the deformation of $\mathcal{O}(t)$. This friction causes the mechanical deformation energy to dissipate into heat, and the length of a path in shape space will be based on the total dissipation along the path.

We will assume the dissipation, induced at some time instant by a motion field v of a purely viscous material, to depend only on the spatial gradient $\mathcal{D}v$ of the velocity field v , which is consistent with the fact that global shifts of the material cause no internal friction and thus no dissipation. Furthermore, since rotational flows are free of friction, too, the viscous dissipation can only depend on the symmetrised gradient

$$\epsilon[v] = \frac{1}{2}(\mathcal{D}v + \mathcal{D}v^T)$$

in which any skew-symmetric components are eliminated. These skew-symmetric components of the gradient describe infinitesimal rotations, as the tangent space to $SO(d)$ at the identity is the vector space of skew-symmetric matrices (and hence, the velocity

fields belonging to rotations have skew-symmetric gradients). If we additionally assume the underlying material to be isotropic and Newtonian (that is, the friction is proportional to the rate of shear), then the viscous dissipation of a velocity field $v : \mathcal{O} \rightarrow \mathbb{R}^d$ is given by

$$g_{\mathcal{O}}(v, v) = \int_{\mathcal{O}} \frac{\lambda}{2} (\operatorname{tr} \epsilon[v])^2 + \mu \operatorname{tr}(\epsilon[v]^2) \, dx$$

for the so-called Lamé parameters λ and μ [78], and the total dissipation along a path $\mathcal{O}(t)$, $t \in [0, 1]$, generated by a velocity field $v(t)$, $t \in [0, 1]$, reads

$$\mathbf{Diss}[v] = \int_0^1 g_{\mathcal{O}(t)}(v(t), v(t)) \, dt.$$

Here, $\operatorname{tr} \epsilon[v(t)] = \operatorname{div} v(t)$ obviously describes the local volume change (indeed, if $\frac{\partial \psi(t, \cdot)}{\partial t} = v(t) \circ \psi(t, \cdot)$ for a flow ψ , then the local rate of volume change is $(\frac{1}{\det \mathcal{D}\psi} \frac{\partial \det \mathcal{D}\psi}{\partial t}) \circ \psi^{-1} = (\frac{\operatorname{cof} \mathcal{D}\psi}{\det \mathcal{D}\psi} : \frac{\partial \mathcal{D}\psi}{\partial t}) \circ \psi^{-1} = (\mathcal{D}\psi^{-T} : [(\mathcal{D}v \circ \psi) \mathcal{D}\psi]) \circ \psi^{-1} = \operatorname{div} v$) so that the first term of $g(v, v)$ represents dissipation due to volume changes, and dissipation due to local length variations enters the equation via $\operatorname{tr}(\epsilon[v]^2) = \|\epsilon[v]\|_F^2$.

Having defined the first fundamental form $g_{\mathcal{O}}(v, v)$ for a tangent vector $v : \mathcal{O} \rightarrow \mathbb{R}^d$ to shape space at the shape $\mathcal{S} = \partial \mathcal{O}$, the Riemannian metric for any two tangent vectors v, \tilde{v} is given by

$$g_{\mathcal{O}}(v, \tilde{v}) = \int_{\mathcal{O}} \frac{\lambda}{2} \operatorname{tr} \epsilon[v] \operatorname{tr} \epsilon[\tilde{v}] + \mu \operatorname{tr}(\epsilon[v] \epsilon[\tilde{v}]) \, dx.$$

A geodesic between two shapes $\mathcal{S}_1 = \partial \mathcal{O}_1$ and $\mathcal{S}_2 = \partial \mathcal{O}_2$ then is a path $\mathcal{S}(t) = \partial \mathcal{O}(t)$, $t \in [0, 1]$, with $\mathcal{S}(0) = \mathcal{S}_1$ and $\mathcal{S}(1) = \mathcal{S}_2$ and a generating motion field $v(t)$, $t \in [0, 1]$, such that $\mathbf{Diss}[v] = \int_0^1 g_{\mathcal{O}(t)}(v(t), v(t)) \, dt$ is minimised.

The physical soundness of the shape space structure is here ensured by the interpretation of shapes as boundaries of objects made of a viscous material and the definition of distance between two shapes via the energy loss associated with an optimal, dissipation-minimising deformation between them. Furthermore, the viscous dissipation provides a very natural Riemannian metric for many applications since it measures the rate of distortion of isometries and prefers paths along which isometries are preserved as much as possible. Since the dissipation of rigid body motions is zero, the geodesic distance between translated or rotated versions of the same shape vanishes so that they are identified with each other. Finally, the viscous approach also allows for a physically sound interpretation of topology changes as for example a closure of a material gap: In this case, the viscous material simply flows into the gap until it is closed (which implies the contact of two material boundaries, compare Figure 6.1).

The process of topology changes and its mathematical representation deserve few more clarifying remarks: The objects \mathcal{O} are modelled as open subsets of \mathbb{R}^d , and the flow generated by a velocity field is continuous and injective. Hence, as a closure of a gap we consider the limit process of two material parts flowing arbitrarily close towards

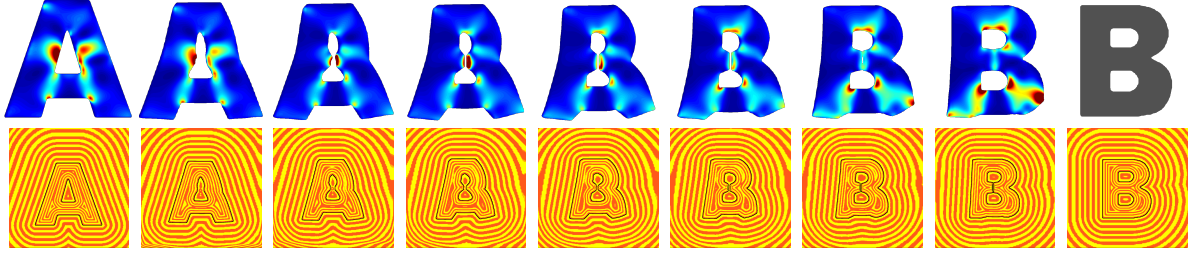


Figure 6.1: Approximate geodesic between the letters A and B . Geodesic distance is measured on the basis of viscous dissipation inside the objects (colour-coded from blue, low dissipation, to red, high dissipation), induced by the motion field which generates the deformation along the path. Topological changes are interpreted as viscous material parts flowing arbitrarily close towards each other. (In the computation, shapes are actually represented via level set functions, whose level lines are texture-coded in the bottom row.)

each other without actually touching each other. Likewise, if material breaks up we suppose that it has an infinitesimally small gap right from the beginning.

In contrast to the concepts of elasticity, viscoelasticity, and viscoplasticity, the dissipation rate and the current internal stress configuration of a purely viscous flow solely depend on the Jacobian of the current velocity field and not on the history of the deformation or the original configuration of the material. In fact, this locality in time is crucial for the definition of a Riemannian metric.

6.2 Variational time discretisation

In order to effectively compute approximations to geodesics between two shapes we have to find a suitable time discretisation of paths and their dissipation. A straightforward linear time discretisation of the motion field v will neither ensure rigid body motion invariance nor a one-to-one correspondence between consecutive shapes.

As a simple example, let us consider a time-discrete path of three shapes $\mathcal{S}_0, \mathcal{S}_1, \mathcal{S}_2$ in shape space and the corresponding objects $\mathcal{O}_0, \mathcal{O}_1, \mathcal{O}_2$, where object $\mathcal{O}_0 = [-1, 1]^2 \subset \mathbb{R}^2$ is the unit square and \mathcal{O}_1 and \mathcal{O}_2 are rotated versions of \mathcal{O}_0 with $\mathcal{O}_1 = \phi_1(\mathcal{O}_0)$ and $\mathcal{O}_2 = \phi_2(\mathcal{O}_1)$ for two rotations

$$\phi_k(x) = \begin{pmatrix} \cos \varphi_k & -\sin \varphi_k \\ \sin \varphi_k & \cos \varphi_k \end{pmatrix} x, \quad k = 1, 2,$$

by angles $\varphi_1, \varphi_2 \in \mathbb{R}$ (Figure 6.2, top). All three shapes are identical up to a rigid body motion so that this time-discrete path should have zero length in shape space. However, letting $\tau > 0$ be the time step size between the shapes, the natural approximation of the

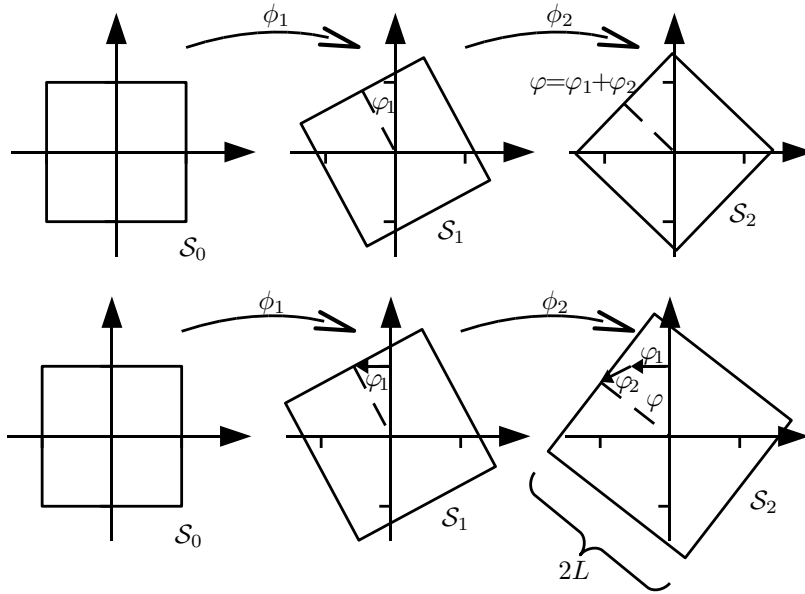


Figure 6.2: Top: Time-discrete approximation of a path in shape space along which \mathcal{S}_0 is rotated by $\varphi_1 + \varphi_2$. Bottom: Time-discrete path in shape space where \mathcal{S}_1 and \mathcal{S}_2 are obtained as linearised rotations of \mathcal{S}_0 and \mathcal{S}_1 , respectively.

corresponding motion field is given by

$$v_\tau^k(x) = \frac{(\phi_k(x) - x)}{\tau} = \frac{1}{\tau} \begin{pmatrix} \cos \varphi_k - 1 & -\sin \varphi_k \\ \sin \varphi_k & \cos \varphi_k - 1 \end{pmatrix} x$$

in the time interval $[(k-1)\tau, k\tau]$. The obvious approximation of the total dissipation then reads

$$\begin{aligned} \text{Diss}_\tau[v_\tau] &= \tau \sum_{k=1}^2 \int_{\mathcal{O}_{k-1}} \frac{\lambda}{2} (\text{tr} \epsilon[v_\tau^k])^2 + \mu \text{tr}(\epsilon[v_\tau^k]^2) dx \\ &= 2(\lambda + \mu) \sum_{k=1}^2 \int_{\mathcal{O}_{k-1}} (\cos \varphi_k - 1)^2 dx = 8(\lambda + \mu) \sum_{k=1}^2 (\cos \varphi_k - 1)^2 \end{aligned}$$

which should actually vanish but is in fact strictly larger than zero due to the lacking rigid body motion invariance of the time discretisation.

We can also observe the opposite phenomenon: Consider a time-discrete geodesic between $\mathcal{O}_0 = [-1, 1]^2$ and $\mathcal{O}_2 = Q([-L, L]^2)$ with one intermediate object \mathcal{O}_1 , where $L > 1$ and Q is a rotation by the angle φ (Figure 6.2, bottom). \mathcal{O}_0 and \mathcal{O}_2 are obviously distinct so that a geodesic between them should have positive length or, equivalently, a continuous deformation of \mathcal{O}_0 into \mathcal{O}_2 should produce dissipation. However, we can find two linearised rotations

$$\phi_k(x) = \begin{pmatrix} 1 & -\varphi_k \\ \varphi_k & 1 \end{pmatrix} x, \quad k = 1, 2,$$

and an object \mathcal{O}_1 with $\phi_1(\mathcal{O}_0) = \mathcal{O}_1$ and $\phi_2(\mathcal{O}_1) = \mathcal{O}_2$ such that $\mathbf{Diss}_\tau[v_\tau] = 0$ (since we chose ϕ_1 and ϕ_2 as linearised rotations). The underlying mechanism is that each linearised rotation by the angle φ_k produces a scaling of $\sqrt{1 + \varphi_k^2}$. Even worse, if the angle φ of the end shape \mathcal{S}_2 is slightly changed, the intermediate angles φ_1 and φ_2 and thus also the size of the intermediate object \mathcal{O}_1 vary. Consequently, the shortest time-discrete path is not independent of the position of the end shapes! (For given L and φ , the two angles φ_1, φ_2 of the linearised rotations can be found as solutions to the nonlinear equation

$$Q(Lx) = \phi_2 \circ \phi_1(x) \quad \Leftrightarrow \quad L \begin{pmatrix} \cos \varphi & -\sin \varphi \\ \sin \varphi & \cos \varphi \end{pmatrix} = \begin{pmatrix} 1 - \varphi_1 \varphi_2 & -\varphi_1 - \varphi_2 \\ \varphi_1 + \varphi_2 & 1 - \varphi_1 \varphi_2 \end{pmatrix}.$$

For example, $L = 2$ and $\varphi = 0$ imply $\varphi_1 = 1, \varphi_2 = -1$, and $L = 2$ and $\varphi = \frac{\pi}{3}$ yield $\varphi_1 = 0, \varphi_2 = \sqrt{3}$.)

The above time discretisation is not the only possible choice, of course, but it serves to illustrate the problems associated with a time discretisation of the motion field v . For an alternative time discretisation to be developed in this section, we pursue the following goals.

- The computation of approximate geodesics shall yield satisfactory results even for a very coarse time discretisation with only few intermediate shapes.
- The time discretisation shall preserve the rigid body motion invariance of the continuous geodesic, and it shall ensure a one-to-one correspondence between consecutive shapes along the path.
- The obtained time-discrete geodesics shall approximate the time-continuous geodesics for a decreasing time step size.
- The discretisation shall facilitate an efficient multiscale solution. More precisely, we first would like to find a geodesic on a coarse time and space resolution to obtain the large scale deformations and then successively refine the results.

Specifically, we will proceed in a variational way and choose a time discretisation which can be regarded as the infinite-dimensional counterpart of the following time discretisation for a geodesic between two fixed points s_A and s_B on a finite-dimensional Riemannian manifold: Consider a sequence of points $s_A = s_0, s_1, \dots, s_K = s_B$ connecting s_A and s_B . To find an approximate geodesic, we minimise

$$\sum_{k=1}^K \text{dist}^2(s_{k-1}, s_k),$$

where $\text{dist}(\cdot, \cdot)$ is a suitable approximation of the Riemannian distance.

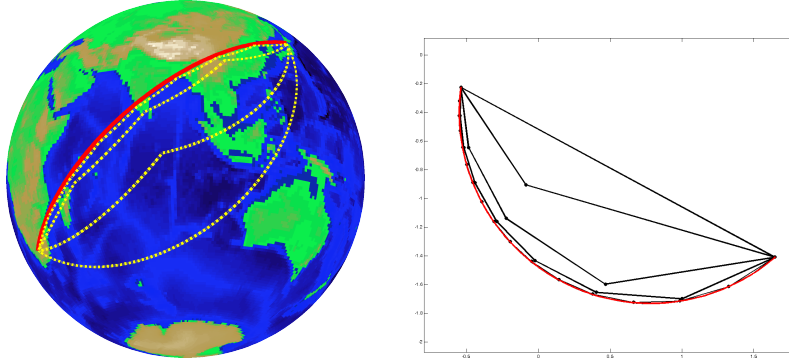


Figure 6.3: Different refinement levels of discrete geodesics ($K = 1, 2, 4, \dots, 256$) from Johannesburg to Kyoto in the stereographic projection (right) and backprojected on the globe (left). A single-level nonlinear Gauss–Seidel iteration on the finest resolution with successive relaxation of the different vertices requires 917235 elementary relaxation steps, whereas in a cascadic relaxation from coarse to fine resolution in time, only 2593 of these elementary minimisation steps are needed.

A conceptual sketch of this approach is depicted in Figure 6.3, where time-discrete geodesics have been computed on the stereographic projection of the two-dimensional sphere. The length $\text{dist}(s_{k-1}, s_k)$ of each segment here is approximated using the Riemannian metric at the segment centre. Obviously, the time-discrete geodesics get the closer to the real geodesic the more points are considered in time.

The need for an efficient minimisation strategy already becomes apparent in this low-dimensional example: A significant speedup can be achieved if first a relaxation on the coarse time resolution is performed and then the number of time points is successively increased.

In our approach, the squared approximate distance $\text{dist}^2(s_{k-1}, s_k)$ will be replaced by the (elastic) deformation energy of a matching deformation between two consecutive shapes \mathcal{S}_{k-1} and \mathcal{S}_k as will be made more precise in the next section.

6.2.1 Discrete geodesics

Given two shapes $\mathcal{S}_A, \mathcal{S}_B$ along with the corresponding objects $\mathcal{O}_A, \mathcal{O}_B$, we define a discrete path of shapes as a sequence of shapes $\mathcal{S}_0, \mathcal{S}_1, \dots, \mathcal{S}_K$ with $\mathcal{S}_0 = \mathcal{S}_A$ and $\mathcal{S}_K = \mathcal{S}_B$. For the time step $\tau = \frac{1}{K}$, the shape \mathcal{S}_k is supposed to be an approximation of $\mathcal{S}(t_k)$, $t_k = k\tau$, where $\mathcal{S}(t)$, $t \in [0, 1]$, is a continuous path connecting $\mathcal{S}_A = \mathcal{S}(0)$ and $\mathcal{S}_B = \mathcal{S}(1)$, for example, a geodesic between these two shapes.

Next, we consider for each pair of consecutive shapes $\mathcal{S}_{k-1}, \mathcal{S}_k$ a matching deformation ϕ_k such that $\phi_k(\mathcal{S}_{k-1}) = \mathcal{S}_k$. This matching deformation shall be associated with a

nonlinear deformation energy

$$\mathcal{W}[\mathcal{O}_{k-1}, \phi_k] = \int_{\mathcal{O}_{k-1}} W(\mathcal{D}\phi_k) \, dx .$$

This deformation energy shall be of the same type as the hyperelastic deformation energies from the previous chapters; however, here it will only serve as an approximation to the viscous, non-elastic dissipation along a path segment. Note that, in fact, we simply encode the desired properties of our time discretisation by this particular choice of a matching energy. In particular, the deformation energy is rigid body motion invariant and ensures injectivity of the matching deformations ϕ_k (under sufficient growth conditions, see Section 3.1) and thus a one-to-one correspondence between consecutive objects \mathcal{O}_{k-1} and \mathcal{O}_k . Furthermore, it is capable of describing large deformations, an essential property if we aim for a coarse time discretisation.

As already discussed in Section 5.2.3, we here assume an instantaneous stress relaxation after each matching deformation ϕ_k , which is fundamentally different from the elasticity approach in the previous two chapters. More precisely, during the deformation ϕ_k , the object \mathcal{O}_{k-1} is regarded as initially tension-free and not as a prestressed, deformed configuration of \mathcal{O}_0 . Also, different from viscoelasticity and viscoplasticity, we shall assume the energy to be independent of the stress history. If the deformation ϕ_k produces a self-contact at the boundary of \mathcal{O}_{k-1} , this will be considered to correspond to a topology change.

We can now define a discrete geodesic, using a variational time discretisation where the Riemannian distance between consecutive points is approximated via the nonlinear deformation energy (we postpone a detailed examination of the relation between time-discrete and time-continuous geodesics to a later section).

Definition 9 (Discrete Geodesic). *A discrete path $\mathcal{S}_0, \mathcal{S}_1, \dots, \mathcal{S}_K$ connecting two shapes \mathcal{S}_A and \mathcal{S}_B is a discrete geodesic if there exists an associated family of deformations ϕ_k with $\phi_k(\mathcal{S}_{k-1}) = \mathcal{S}_k$ which minimises the total deformation energy $\sum_{k=1}^K \mathcal{W}[\mathcal{O}_{k-1}, \phi_k]$.*

Figure 6.4 proves that both the built-in exact frame indifference and the one-to-one mapping property ensure that fairly coarse time discretisations already lead to an accurate approximation of geodesic paths. Furthermore, isometries are locally preserved by the nonlinear deformation energy as is illustrated in Figure 6.5.

The definition of a discrete geodesic in its above form is not suited to allow for topological changes in the sense that we have described in the previous section: If the boundary of a deformed object $\phi_k(\mathcal{O}_{k-1})$ exhibits self-contact, which represents a topological transition between $\phi_k(\mathcal{O}_{k-1})$ and $\overline{\mathcal{O}_k}$, then $\phi_k(\mathcal{S}_{k-1})$ will exhibit an interior edge along the contact line while \mathcal{S}_k has no edge there. The same holds true for \mathcal{S}_{k-1} and $\phi_k^{-1}(\mathcal{S}_k)$ for topological transitions in the opposite direction. More precisely, since the connecting deformations ϕ_k are supposed to be continuous and injective, the constraints imply that all shapes \mathcal{S}_k are homeomorphic. In order to enable topological transitions, the constraints

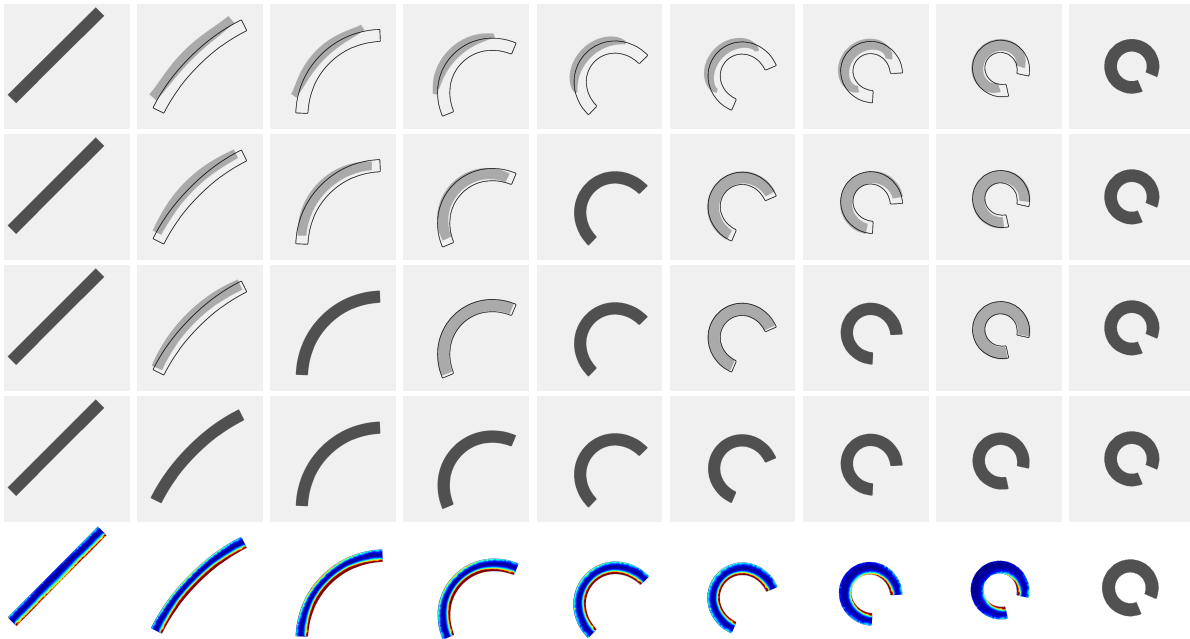



Figure 6.4: Discrete geodesics between a straight and a rolled up bar, based on 1, 2, 4, and 8 time steps (dark grey shapes in first to fourth row). The light grey shapes in each row show the linear interpolation of the deformations connecting the dark grey shapes. The shapes from the finest time discretisation are overlaid over the others as thin black contour lines. In the last row, the rate of viscous dissipation is rendered on the shape domains $\mathcal{O}_1, \dots, \mathcal{O}_{K-1}$ from the previous row, colour-coded as .

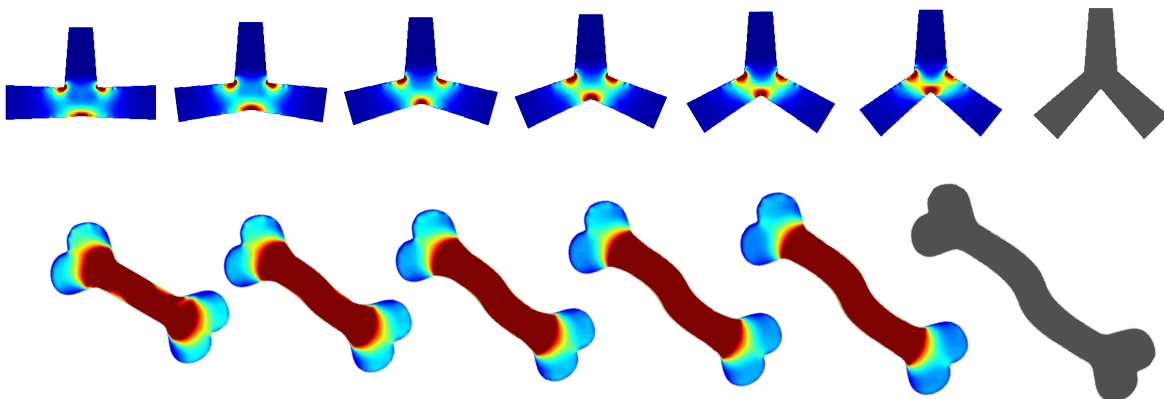



Figure 6.5: Discrete geodesic for two different examples from [30] and [62] where the local rate of dissipation is colour-coded as . The local preservation of isometries is clearly visible in both examples.

$\phi_k(\mathcal{S}_{k-1}) = \mathcal{S}_k$ have to be replaced by the condition $\overline{\phi_k(\mathcal{O}_{k-1})} = \overline{\mathcal{O}_k}$ for $k = 1, \dots, K$, which will allow a material breakup as well as the gap-filling flow from the previous section. In this case, however, we certainly have to restrict the admissible set of shapes, for example, by adding for each intermediate shape \mathcal{S}_k a regularisation energy of the form

$$\mathcal{L}[\mathcal{S}] = \mathcal{H}^{d-1}(\mathcal{S})$$

to the total dissipation. Otherwise it would be optimal to decompose the initial object \mathcal{O}_0 into tiny pieces, shuffle these around, and remerge them to obtain the final object \mathcal{O}_K . Figure 6.1 has already proven the feasibility of this approach to describe topological changes along a path in shape space.

6.2.2 A relaxed formulation

The constraints $\phi_k(\mathcal{S}_{k-1}) = \mathcal{S}_k$ and $\overline{\phi_k(\mathcal{O}_{k-1})} = \overline{\mathcal{O}_k}$ are inconvenient with respect to numerical treatment, and they are not robust with respect to noise in the shape acquisition process. As in the previous chapters, we will replace the constraint by adding a mismatch penalty to the energy in the definition of a discrete geodesic. As explained in the previous section, only a volumetric penalty is reasonable in the context of (cost-free) topological changes since self-contact of $\phi_k(\mathcal{O}_{k-1})$ along line segments of non-vanishing length would be inhibited by edge-based mismatch penalties. Also, a volumetric measure of the mismatch is less prone to confusing different regions of a shape while an edge-based approach may for example easily match two neighbouring edges of a shape. The robustness of the mismatch penalty is especially important in a sequence of matching problems (as we have here), where incorrect matches are passed on and errors accumulate along the sequence. We hence deliberately use a volumetric mismatch penalty as opposed to the edge-based penalty from Chapter 4.

This mismatch penalty is chosen as the Lebesgue measure of the symmetric difference between \mathcal{O}_{k-1} and the preimage of \mathcal{O}_k under ϕ_k , $\phi_k^{-1}(\mathcal{O}_k)$,

$$\mathcal{F}[\mathcal{O}_{k-1}, \phi_k, \mathcal{O}_k] = |\mathcal{O}_{k-1} \Delta \phi_k^{-1}(\mathcal{O}_k)| := |(\mathcal{O}_{k-1} \setminus \phi_k^{-1}(\mathcal{O}_k)) \cup (\phi_k^{-1}(\mathcal{O}_k) \setminus \mathcal{O}_{k-1})|.$$

As already mentioned, one additionally has to restrict the set of admissible shapes \mathcal{S}_k along a discrete geodesic by adding an extra surface energy term

$$\mathcal{L}[\mathcal{S}_k] = \mathcal{H}^{d-1}(\mathcal{S}_k)$$

that prevents arbitrarily irregular shape boundaries. In particular, such a regulariser will reduce the formation of cracks (see Figure 6.6) that are associated with a local change of topology and can therefore easily appear along a geodesic path where topological transitions cost no additional energy.

One may also be interested in an underlying shape space of shapes enclosing a constant volume V , $|\mathcal{O}_k| \equiv V$ (compare Figure 6.6, bottom). In this case we will add a further

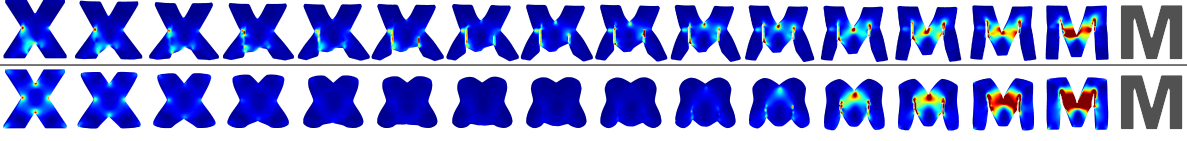


Figure 6.6: Geodesic paths between an X and an M without a contour length term ($\eta = 0$, top row), allowing for crack formation, and with this term damping down cracks and rounding corners (bottom row). In the bottom row we additionally enforce area preservation along the geodesic.

penalty term of volume deviation,

$$\mathcal{V}[\mathcal{O}_k] = (|\mathcal{O}_k| - V)^2.$$

Finally, we end up with the total discrete energy

$$\mathcal{E}_\tau[(\phi_k, \mathcal{S}_{k-1}, \mathcal{S}_k)_{k=1, \dots, K}] = \sum_{k=1}^K \left(\frac{1}{\tau} \mathcal{W}[\mathcal{O}_{k-1}, \phi_k] + \gamma \mathcal{F}[\mathcal{O}_{k-1}, \phi_k, \mathcal{O}_k] + \eta \tau \mathcal{L}[\mathcal{S}_k] + \nu \tau \mathcal{V}[\mathcal{O}_k] \right)$$

where γ , η , ν are positive weights, and a minimiser of this energy for fixed shapes \mathcal{S}_0 and \mathcal{S}_K describes a *relaxed discrete geodesic path* between \mathcal{S}_0 and \mathcal{S}_K . The scaling of the different terms by the time step size τ ensures that none of the energy contributions converge to zero or infinity as $\tau \rightarrow 0$, as will become clear in the next section.

6.2.3 Viscous fluid model for vanishing time step size

Next, we will investigate the relation between the above-introduced relaxed discrete geodesic paths and the time-continuous model for geodesics in shape space.

Given a sequence $\mathcal{S}_0, \dots, \mathcal{S}_K$ of shapes and corresponding, approximately connecting deformations ϕ_1, \dots, ϕ_K , we can define a time-continuous, interpolating flow ψ_τ and path \mathcal{S}_τ in shape space by

$$\begin{aligned} v_\tau^k &:= \frac{1}{\tau}(\phi_k - \text{id}), \\ \phi_\tau^k(t) &:= (\text{id} + (t - t_{k-1})v_\tau^k), \\ \psi_\tau(t) &:= \phi_\tau^k(t) \circ \phi_{k-1} \circ \dots \circ \phi_1, \\ \mathcal{S}_\tau(t) &:= \phi_\tau^k(t)(\mathcal{S}_{k-1}) \end{aligned}$$

for $t \in [t_{k-1}, t_k)$ with $t_k = k\tau$. The corresponding motion field, which generates the flow, is then given by

$$v_\tau(t) := v_\tau^k \circ \phi_\tau^k(t)^{-1}$$

on $[t_{k-1}, t_k)$, where we assume $\phi_\tau^k(t)$ to be injective, and the concatenation with its inverse is only needed to obtain the proper Eulerian description of the motion field.

If we now let $\tau \rightarrow 0$ and assume that $\mathcal{S}_\tau(t) \rightarrow \mathcal{S}(t)$ for a regular family of shapes $(\mathcal{S}(t))_{0 \leq t \leq 1}$ and that $v_\tau(t) \rightarrow v(t)$ with $\frac{\partial \psi(t, \cdot)}{\partial t} \circ \psi(t, \cdot)^{-1} = v(t)$ for a sufficiently regular flow $\psi : [0, 1] \times \mathcal{O}(0) \rightarrow \mathbb{R}^d$, the following limit behaviour can be observed: The first term in the time-discrete energy \mathcal{E}_τ , the sum $\sum_{k=1}^K \frac{1}{\tau} \mathcal{W}[\mathcal{O}_{k-1}, \phi_k]$ of deformation energies, turns into the time-continuous dissipation functional

$$\mathbf{Diss}[v] = \int_0^1 \int_{\mathcal{O}(t)} \frac{1}{2} \mathbf{C} \epsilon[v(t)] : \epsilon[v(t)] \, dx \, dt,$$

where the tensor $\mathbf{C} = W_{,AA}(I)$ is the Hessian of the energy density at the identity and $\epsilon[v] = \frac{1}{2}(\mathcal{D}v^T + \mathcal{D}v)$ denotes the symmetrised gradient of the motion field. Indeed, by second order Taylor expansion about the identity we observe

$$W(\mathcal{D}\phi_k) = W(I) + \tau W_{,A}(I) : \mathcal{D}v_\tau^k + \frac{\tau^2}{2} W_{,AA}(I) \mathcal{D}v_\tau^k : \mathcal{D}v_\tau^k + O(\tau^3) = \frac{\tau^2}{2} \mathbf{C} \mathcal{D}v_\tau^k : \mathcal{D}v_\tau^k + O(\tau^3),$$

noting that, at the identity, the energy density W attains its minimum 0 and the first Piola–Kirchhoff stress $\sigma^{\text{ref}} = W_{,A}(I)$ vanishes (Section 3.1). Furthermore, the rigid body motion invariance of W implies $\mathbf{C}A = 0$ for all skew-symmetric matrices $A \in \mathbb{R}^{d \times d}$ and thus $\mathbf{C} \mathcal{D}v : \mathcal{D}v = \mathbf{C} \epsilon[\mathcal{D}v] : \epsilon[\mathcal{D}v]$. We may now choose the deformation energy density in such a way that we obtain

$$\frac{1}{2} \mathbf{C} \epsilon[v] : \epsilon[v] = \frac{\lambda}{2} (\text{tr} \epsilon[v])^2 + \mu \text{tr}(\epsilon[v]^2)$$

for given parameters λ and μ . In the limit $\tau \rightarrow 0$ we thus recover the desired dissipation integral, and the resulting Riemannian metric, given by the dissipation rate, is associated with the Hessian of the nonlinear deformation energy.

The sum of mismatch penalties, $\sum_{k=1}^K \mathcal{F}[\mathcal{O}_{k-1}, \phi_k, \mathcal{O}_k]$, converges against an optical flow type energy (analogous to L^1 -type optical flow functionals like in [16]),

$$\mathcal{E}_{\text{OF}}[v, \mathcal{S}] = \int_{\mathcal{T}} |(1, v(t))^T \cdot \nu[\mathcal{T}]| \, da,$$

where $\mathcal{T} = \bigcup_{t \in [0, 1]} (t, \mathcal{S}(t)) \subset [0, 1] \times \mathbb{R}^d$ denotes the shape tube in space-time, $(1, v(t))$ is the underlying space-time motion field, and $\nu[\mathcal{T}]$ the unit outward normal to \mathcal{T} . (Note that this integral can also be rewritten in the intuitive, non-rigorous, classical optical flow form

$$\mathcal{E}_{\text{OF}}[v, \mathcal{S}] = \int_0^1 \int_{\mathbb{R}^d} \left| \partial_t \chi_{\mathcal{O}(t)} - v(t) \cdot \nabla_x \chi_{\mathcal{O}(t)} \right| \, dx \, dt,$$

where $\chi_{\mathcal{O}(t)}$ denotes the characteristic function of $\mathcal{O}(t)$ in $SBV(\mathbb{R}^d)$.) To see this, let us regard $|\mathcal{O}_{k-1} \Delta \phi_k^{-1}(\mathcal{O}_k)|$ as the mismatch induced within time τ by the motion field v_τ^k , which is not consistent with the correct flow between \mathcal{S}_{k-1} and \mathcal{S}_k . A correct flow $(1, \hat{v})$

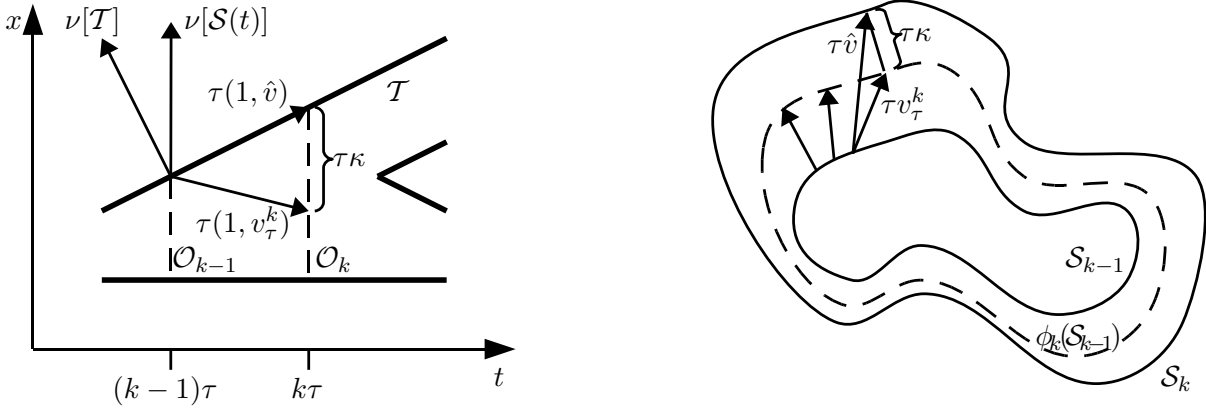


Figure 6.7: Sketch of the mismatch between shapes and motion fields, for one-dimensional shapes on the left and for two-dimensional shapes on the right. Note that the mismatch κ is visualised between $\phi_k(\mathcal{S}_{k-1})$ and \mathcal{S}_k for better understandability, but it is actually measured on \mathcal{S}_{k-1} as the mismatch between \mathcal{S}_{k-1} and $\phi_k^{-1}(\mathcal{S}_k)$.

can be obtained if the motion vector $(1, v_\tau^k)$ is projected back onto the tube of shapes \mathcal{T} along the normal $\nu[\mathcal{S}_{k-1}]$ to \mathcal{S}_{k-1} (Figure 6.7). This leads to the equations

$$(1, \hat{v})^\top \cdot \nu[\mathcal{T}] = 0 \quad \text{and} \quad (1, \hat{v}) = (1, v_\tau^k) + \kappa \nu[\mathcal{S}_{k-1}]$$

for some $\kappa \in \mathbb{R}$, and upon substitution we have $(1, v_\tau^k)^\top \cdot \nu[\mathcal{T}] = -\kappa \nu[\mathcal{S}_{k-1}] \cdot \nu[\mathcal{T}]$. The local rate at which $\phi_\tau^k(t)(\mathcal{S}_{k-1})$ and the tube of shapes \mathcal{T} diverge on the time interval $[t_{k-1}, t_k]$ is then given as $|(1, \hat{v}) - (1, v_\tau^k)| = |\kappa| = \left| (1, v_\tau^k(t))^\top \cdot \frac{\nu[\mathcal{T}]}{\nu[\mathcal{T}] \cdot \nu[\mathcal{S}_{k-1}]} \right|$. Hence, we obtain

$$\mathcal{E}_{\text{OF}}[v, \mathcal{S}] = \int_0^1 \int_{\mathcal{S}(t)} \left| (1, v(t))^\top \cdot \frac{\nu[\mathcal{T}]}{\nu[\mathcal{T}] \cdot \nu[\mathcal{S}(t)]} \right| da dt,$$

which turns into the desired integral by a change of variables according to $d\mathcal{H}^{d-1} \llcorner \mathcal{S} dt = \nu[\mathcal{T}] \cdot \nu[\mathcal{S}(t)] d\mathcal{H}^d \llcorner \mathcal{T}$.

Finally, the sum of shape perimeters, $\sum_{k=1}^K \tau \mathcal{L}[\mathcal{S}_k]$, of course converges against the time integral of the perimeters,

$$\int_0^1 \mathcal{L}[\mathcal{S}(t)] dt,$$

and the volume mismatch, $\sum_{k=1}^K \tau \mathcal{V}[\mathcal{O}_k]$, turns into

$$\int_0^1 \mathcal{V}[\mathcal{O}(t)] dt$$

so that the total limit energy reads

$$\mathcal{E}[v, \mathcal{S}] = \mathbf{Diss}[v] + \gamma \mathcal{E}_{\text{OF}}[v, \mathcal{S}] + \eta \int_0^1 \mathcal{L}[\mathcal{S}(t)] dt + \nu \int_0^1 \mathcal{V}[\mathcal{O}(t)] dt.$$

We may conclude that our variational time discretisation is indeed consistent with the time-continuous viscous dissipation model of geodesic paths. The total dissipation along a path and the corresponding geodesic length are approximated by

$$\sum_{k=1}^K \frac{1}{\tau} \mathcal{W}[\mathcal{O}_{k-1}, \phi_k] \quad \text{and} \quad \sum_{k=1}^K \sqrt{\mathcal{W}[\mathcal{O}_{k-1}, \phi_k]},$$

respectively. Note that we are interested in the case $\gamma \gg 1$ since the L^1 -type optical flow term is supposed to just act as a penalty, whereas the L^2 -type counterpart may actually be used as part of a geodesic distance [87, 88].

6.3 Regularised level set approximation

We have already discussed in Section 1.2 that—concerning the spatial discretisation of the time-discrete energy—an explicit triangulation of the objects \mathcal{O}_k is associated with a number of problems. The intermediate shapes are unknown a priori and thus cannot easily be triangulated. If the triangulation is transported along the path, then it might degenerate (and it will certainly in the case of topological transitions). Also, a spatial multiscale approach will be difficult to implement.

Instead, we choose to describe the objects $\mathcal{O}_0, \dots, \mathcal{O}_K$ implicitly as the zero super-level set of level set functions $u_0, \dots, u_K : \Omega \rightarrow \mathbb{R}$, defined on some computational domain $\Omega \subset \mathbb{R}^d$, according to

$$\mathcal{O}_k = \{x \in \Omega : u_k(x) > 0\}.$$

Unlike a phase field description, this approach is less prone to artificial anisotropies that can be observed experimentally for phase fields: In order to minimise the phase field interface energy, the interfaces align to the grid which allows them to attain an almost optimal phase field profile. This alignment (at least for coarse spatial discretisations) is not negligible when computing geodesics via the above time discretisation since the fixed first and last shape, \mathcal{S}_0 and \mathcal{S}_K , act as direct shape priors only on the neighbouring shapes. Hence, their influence decreases towards the middle of the geodesic so that the grid alignment is hardly counteracted there.

As in the previous chapters, we will extend the deformations ϕ_k to be defined on the entire domain Ω . To regularise ϕ_k outside \mathcal{O}_{k-1} , the region $\Omega \setminus \mathcal{O}_{k-1}$ is also regarded as viscous with the material parameters reduced by the factor $\delta = 10^{-4}$.

We will approximate the characteristic function $\chi_{\mathcal{O}_k}$ of object \mathcal{O}_k as a concatenation of the level set function u_k with the regularised Heaviside function

$$H_\varepsilon(x) = \frac{1}{2} + \frac{1}{\pi} \arctan\left(\frac{x}{\varepsilon}\right)$$

from [27], where the scale parameter ε represents the width of the smeared transition region (compare Section 2.3). With respect to the numerical minimisation of the energy,

the nonlocal support of the derivative H'_ε is crucial since it allows the zero level set to be guided by distant features. Within the level set setting, the different energy terms have to be restated as

$$\begin{aligned}\mathcal{W}^\varepsilon[u_{k-1}, \phi_k] &= \int_{\Omega} ((1 - \delta)H_\varepsilon(u_{k-1}) + \delta) W(\mathcal{D}\phi_k) \, dx, \\ \mathcal{F}^\varepsilon[u_{k-1}, \phi_k, u_k] &= \int_{\Omega} (H_\varepsilon(u_k(\phi_k)) - H_\varepsilon(u_{k-1}))^2 \, dx, \\ \mathcal{L}^\varepsilon[u_k] &= \int_{\Omega} |\nabla H_\varepsilon(u_k)| \, dx, \\ \mathcal{V}^\varepsilon[u_k] &= \left(\int_{\Omega} H_\varepsilon(u_k) \, dx - V \right)^2,\end{aligned}$$

where the mismatch penalty has turned into the squared L^2 -difference between the approximate characteristic functions $H_\varepsilon \circ u_k \circ \phi_k$ and $H_\varepsilon \circ u_{k-1}$ and the shape perimeter of \mathcal{S}_k has been replaced by the total variation of $H_\varepsilon \circ u_k$. For the mismatch penalty to be properly defined, we here assume the level set functions to be extended smoothly outside Ω . The total, time-discrete energy finally reads

$$\mathcal{E}_\tau^\varepsilon[(\phi_k, u_{k-1}, u_k)_{k=1, \dots, K}] = \sum_{k=1}^K \left(\frac{1}{\tau} \mathcal{W}^\varepsilon[u_{k-1}, \phi_k] + \gamma \mathcal{F}^\varepsilon[u_{k-1}, \phi_k, u_k] + \eta \tau \mathcal{L}^\varepsilon[u_k] + \nu \tau \mathcal{V}^\varepsilon[u_k] \right).$$

The corresponding Euler–Lagrange equations are obtained by straightforward differentiation. Let us denote by $\langle \delta_z \mathcal{G}, \zeta \rangle$ the variation of an energy \mathcal{G} with respect to the function z in a direction ζ . For sufficiently smooth u_k and ϕ_k we have

$$\begin{aligned}\langle \delta_{\phi_k} \mathcal{W}^\varepsilon[u_{k-1}, \phi_k], \theta \rangle &= \int_{\Omega} ((1 - \delta)H_\varepsilon(u_{k-1}) + \delta) W_{,A}(\mathcal{D}\phi_k) : \mathcal{D}\theta \, dx, \\ \langle \delta_{u_{k-1}} \mathcal{W}^\varepsilon[u_{k-1}, \phi_k], \vartheta \rangle &= \int_{\Omega} (1 - \delta) W(\mathcal{D}\phi_k) H'_\varepsilon(u_{k-1}) \vartheta \, dx, \\ \langle \delta_{\phi_k} \mathcal{F}^\varepsilon[u_{k-1}, \phi_k, u_k], \theta \rangle &= 2 \int_{\Omega} (H_\varepsilon(u_k \circ \phi_k) - H_\varepsilon(u_{k-1})) H'_\varepsilon(u_k \circ \phi_k) \nabla u_k \circ \phi_k \cdot \theta \, dx, \\ \langle \delta_{u_{k-1}} \mathcal{F}^\varepsilon[u_{k-1}, \phi_k, u_k], \vartheta \rangle &= -2 \int_{\Omega} (H_\varepsilon(u_k \circ \phi_k) - H_\varepsilon(u_{k-1})) H'_\varepsilon(u_{k-1}) \vartheta \, dx, \\ \langle \delta_{u_k} \mathcal{F}^\varepsilon[u_{k-1}, \phi_k, u_k], \vartheta \rangle &= 2 \int_{\Omega} (H_\varepsilon(u_k \circ \phi_k) - H_\varepsilon(u_{k-1})) H'_\varepsilon(u_k \circ \phi_k) \vartheta \circ \phi_k \, dx, \\ \langle \delta_{u_k} \mathcal{L}^\varepsilon[u_k], \vartheta \rangle &= 2 \int_{\Omega} H''_\varepsilon(u_k) |\nabla u_k| \vartheta + H'_\varepsilon(u_k) \frac{\nabla u_k \cdot \nabla \vartheta}{|\nabla u_k|} \, dx, \\ \langle \delta_{u_k} \mathcal{V}^\varepsilon[u_k], \vartheta \rangle &= 2 \left(\int_{\Omega} H_\varepsilon(u_k) \, dx - V \right) \int_{\Omega} H'_\varepsilon(u_k) \vartheta \, dx\end{aligned}$$

for scalar-valued test functions ϑ and test displacements θ . The optimality conditions for u_k and ϕ_k , $k = 1, \dots, K$, are then $0 = \langle \delta_{u_i} \mathcal{E}_\tau^\varepsilon[(\phi_k, u_{k-1}, u_k)_{k=1, \dots, K}], \vartheta \rangle$ for $i = 1, \dots, K - 1$

and $0 = \langle \delta_{\phi_i} \mathcal{E}_\tau^\varepsilon[(\phi_k, u_{k-1}, u_k)_{k=1, \dots, K}], \theta \rangle$ for $i = 1, \dots, K$ where $\delta_{u_i} \mathcal{E}_\tau^\varepsilon$ and $\delta_{\phi_i} \mathcal{E}_\tau^\varepsilon$ can be expressed as a sum of the variations of the corresponding energy components.

6.4 Numerical implementation

For the spatial discretisation of the energy $\mathcal{E}_\tau^\varepsilon$ we again employ continuous, multilinear finite elements on a regular, rectangular grid with grid size h , covering the computational domain $\Omega = [0, 1]^d$. The level set functions u_0, \dots, u_K and deformations ϕ_1, \dots, ϕ_K are approximated by finite element functions U_0, \dots, U_K and Φ_1, \dots, Φ_K as described in detail in Section 4.3.

For fixed h and τ , we perform a gradient descent minimisation of $\mathcal{E}_\tau^\varepsilon$ for U_0, \dots, U_K and Φ_1, \dots, Φ_K in the space of continuous multilinear finite element functions. For this purpose, we approximate $\mathcal{E}_\tau^\varepsilon$ numerically, utilising Gaussian quadrature of third order on each grid cell. Furthermore, we compute its Gâteaux derivative with respect to a level set function in the direction of all finite element basis functions φ_i , $(\langle \delta_{u_k} \mathcal{E}_\tau^\varepsilon, \varphi_i \rangle)_{i \in I_h}$, where I_h denotes the index set of all vertices and where we use the formulae for the variation of the different energy components from the previous section. The resulting vector then is the discrete derivative of $\mathcal{E}_\tau^\varepsilon$ with respect to the nodal values of the discrete level set function U_k , and it is used as the descent direction. For the displacements we proceed analogously.

The step size of the gradient descent is determined by Armijo's rule (compare Section 4.3). In fact, we alternate between one joint descent step for all level set functions and one for all deformations (which corresponds to an alternating minimisation scheme). This has the advantage that the backtracking algorithm along the descent direction operates jointly on all level set functions and on all deformations, respectively, which reduces the computational effort as opposed to seeking an optimal step length for each single level set function and deformation. Also, if each level set function and deformation was updated separately, the updating order would have to be chosen very carefully to avoid a biased flow of information along the discrete geodesic. Furthermore, this technique of updating all level set functions and deformations separately, which corresponds to a local, nonlinear Gauss–Seidel smoothing, is experimentally observed to be outperformed by the simultaneous relaxation with respect to the whole set of discrete deformations and discrete level set functions. A separate descent step for the level set functions and the deformations becomes necessary since their step lengths in general differ so that a joint step size control would hamper the descent.

The numerical evaluation of $\mathcal{E}_\tau^\varepsilon[(\Phi_k, U_{k-1}, U_k)_{k=1, \dots, K}]$ and its variations requires the computation of pullbacks $U_k \circ \Phi_k$. If $\Phi_k(x)$ lies inside Ω for a quadrature point x , then the pullback is evaluated exactly at x in our scheme. Otherwise, we project $\Phi_k(x)$ back onto the boundary of Ω and evaluate U_k at that projection point. This procedure is important for two reasons: First, if we only integrated in regions for which $\Phi_k(x) \in \Omega$, we would induce a tendency for Φ_k to shift the domain outwards until $\Phi_k(\Omega) \cap \Omega = \emptyset$

since this would yield zero mismatch penalty. Second, for a gradient descent to work properly, we need a smooth transition of the energy if a quadrature point is displaced outside Ω or comes back in. By the form of the mismatch penalty, this implies that the U_k have to be extended continuously outside Ω . Backprojecting $\Phi_k(x)$ onto the boundary just emulates a constant extension of U_k perpendicular to the boundary.

Concerning the assembly of $(\langle \delta_{u_k} \mathcal{E}_\tau^\varepsilon, \varphi_i \rangle)_{i \in I_h}$, we also have to evaluate pullbacks $\varphi_i \circ \Phi_k$ of finite element basis functions. These are treated as described in Section 5.3, meaning that they are also exactly evaluated (after possibly projecting $\Phi_k(x)$ back onto $\partial\Omega$). If the transformation rule were applied instead, thereby removing the pullbacks, the numerical approximation of the derivative $(\langle \delta_{u_k} \mathcal{E}_\tau^\varepsilon, \varphi_i \rangle)_{i \in I_h}$ would no longer fit the discrete energy exactly, which strongly hampers the gradient descent.

Due to the high nonlinearity of $\mathcal{E}_\tau^\varepsilon$, we expect a very slow convergence of the minimisation for fixed h and τ since information flow in space and time scales with these discretisation parameters. Furthermore, we would like to prevent the algorithm from getting trapped in a nearby local minimum. For this reason, we apply a cascadic multiscale approach: We start the relaxation at a coarse discretisation to obtain the large scale geodesic flow and then successively refine the result. With respect to the spatial discretisation, we employ the same hierarchy of grids with dyadic resolution as introduced in Section 4.3. The regularisation parameter ε is here again coupled to the grid size according to $\varepsilon = h$. Concerning the refinement in time, we add one level set function $U_{k-\frac{1}{2}}$ and deformation $\Phi_{k-\frac{1}{2}}$ in between each two consecutive level set functions U_{k-1} and U_k and initialise $\Phi_{k-\frac{1}{2}}$ as the linear interpolation of the flow associated with Φ_k , $\Phi_{k-\frac{1}{2}} = \text{id} + \frac{1}{2}(\Phi_k - \text{id})$. The level set function $U_{k-\frac{1}{2}}$ is correspondingly initialised as $U_{k-1} \circ \Phi_{k-\frac{1}{2}}^{-1}$, and the minimisation is continued with the finer time discretisation.

The entire algorithm in pseudo code notation reads as follows (where bold capitals represent vectors of nodal values and the $2^j + 1$ shapes on time level j are labelled with the superscript j):

```

EnergyRelaxation ( $U_{\text{start}}, U_{\text{end}}$ ) {
  for time level  $j = j_0$  to  $J$  {
     $K = 2^j$ ;  $U_0^j = U_{\text{start}}$ ;  $U_K^j = U_{\text{end}}$ 
    if ( $j = j_0$ ) {
      initialize  $\Phi_i^j = \text{id}$ ,  $U_i^j = U_K^j$ ,  $i = 1, \dots, K$ 
    } else {
      initialize  $\Phi_{2i-1}^j = \text{id} + \frac{1}{2}(\Phi_i^{j-1} - \text{id})$ ,  $\Phi_{2i}^j = \Phi_i^{j-1} \circ (\Phi_{2i-1}^j)^{-1}$ ,
         $U_{2i}^j = U_i^{j-1}$ ,  $U_{2i-1}^j = U_i^{j-1} \circ \Phi_{2i}^j$ ,  $i = 1, \dots, \frac{K}{2}$ ;
    }
    restrict  $U_i^j$ ,  $\Phi_i^j$  for all  $i = 1, \dots, K$  onto the coarsest grid level  $l_0$ ;
    for grid level  $l = l_0$  to  $L$  {
      for step  $k = 0$  to  $k_{\text{max}}$  {
        perform a gradient descent step
      }
    }
  }
}
    
```

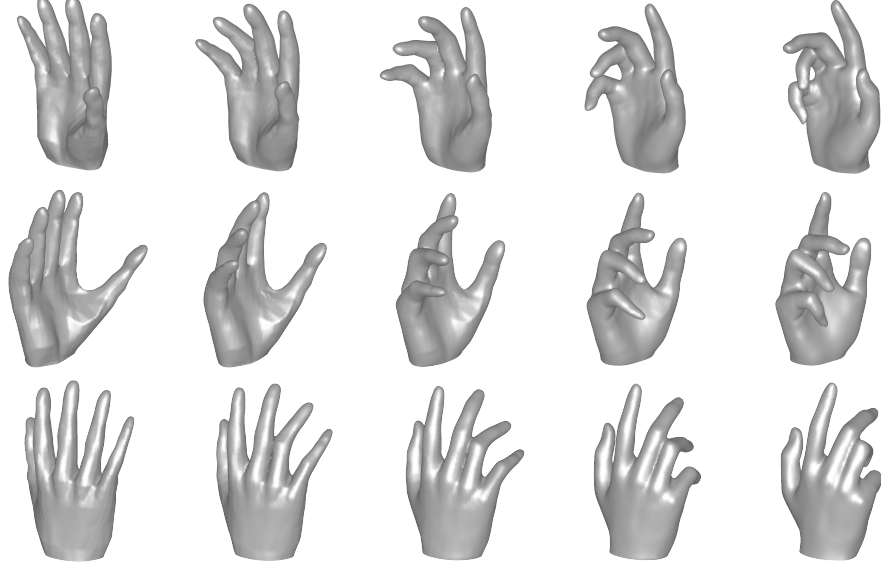


Figure 6.8: Discrete geodesic between the hand shapes *m336* and *m324* from the Princeton Shape Benchmark [108] (the different rows show different views).

$$\begin{aligned}
 & (\Phi_i^j)_{i=1,\dots,K} = (\Phi_i^{j,\text{old}})_{i=1,\dots,K} - \tau \left(\frac{d}{d\Phi_i^j} \mathcal{E}_\tau^\varepsilon[(\Phi_k^j, U_{k-1}^j, U_k^j)_{k=1,\dots,K}] \right)_{i=1,\dots,K} \\
 & \text{with Armijo step size control for } \tau; \\
 & \text{perform a gradient descent step} \\
 & (\mathbf{U}_i^j)_{i=1,\dots,K} = (\mathbf{U}_i^{j,\text{old}})_{i=1,\dots,K} - \tau \left(\frac{d}{d\mathbf{U}_i^j} \mathcal{E}_\tau^\varepsilon[(\Phi_k^j, U_{k-1}^j, U_k^j)_{k=1,\dots,K}] \right)_{i=1,\dots,K} \\
 & \text{with Armijo step size control for } \tau; \\
 & \quad \quad \quad \} \\
 & \quad \quad \text{if } (l < L) \text{ prolongate } U_i^j, \Phi_i^j \text{ for all } i = 1, \dots, K \text{ onto the next grid level;} \\
 & \quad \quad \} \\
 & \quad \quad \} \\
 & \quad \quad \}
 \end{aligned}$$

On a 3 GHz Pentium 4, 2D computations for $K = 8$ and a spatial resolution of 257^2 nodes require (without runtime optimisation) approximately one hour. Since the evaluation of the different energy components can be parallelised, the computation time is downscaled by the factor K in a parallelised implementation.

6.5 Examples and generalisations

We have already seen that the proposed variational time discretisation of geodesics allows to use a very coarse resolution in time (Figure 6.4). Also, the viscous fluid modelling and

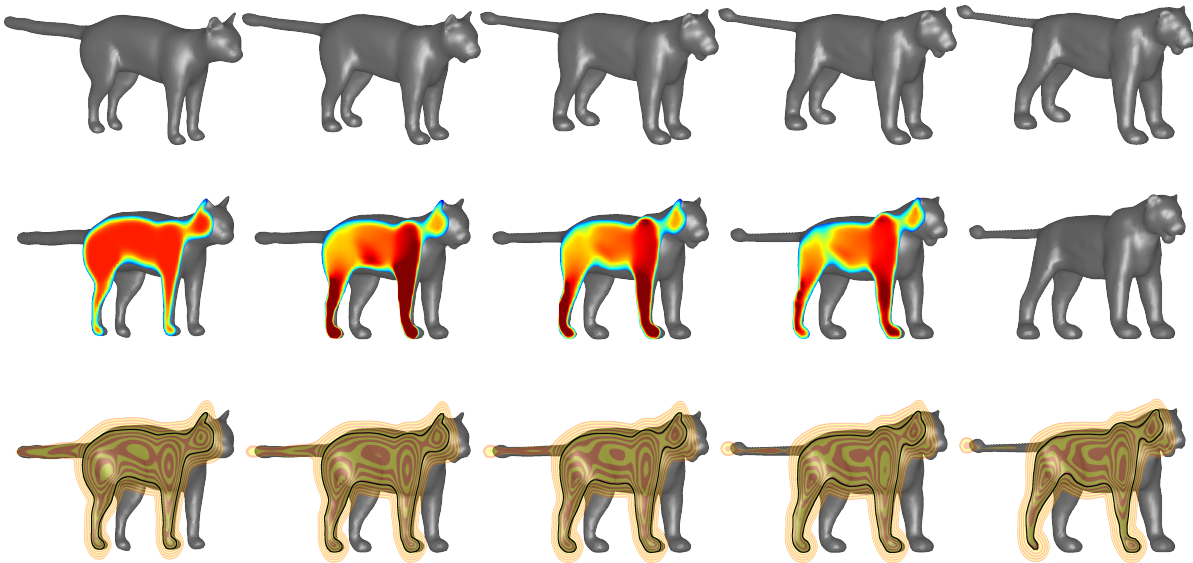



Figure 6.9: Geodesic path between a cat and a lion, with the local rate of dissipation colour-coded as  on the shape interior $\mathcal{O}_0, \dots, \mathcal{O}_{K-1}$ (middle row) and a transparent slicing plane with texture-coded level lines (bottom row).

the representation of shapes via level set functions allows topological changes along the geodesic (Figure 6.1). In particular, unlike in [76], we do not have to equip the starting and the end shape with topologically equivalent meshes. Of course, the volumetric interpretation of shapes comes at the cost of a much higher computational effort already for moderate resolutions. Nevertheless, it is possible to compute geodesics between comparably fine structures (Figure 6.8). A 3D example also computed in [76] is the morphing between a cat and a lion shape, depicted in Figure 6.9. It underlines once more the difference of our approach to a purely geometric view of shapes as manifolds of codimension one, since the viscous dissipation along the geodesic is distributed all over the interior volume (middle row).

6.5.1 A fragment of shape space

In this section, we intend to give an impression of the huge complexity of the Riemannian shape space, which is already revealed by a small example. Figure 6.10 shows a close-up of that part of shape space which is spanned by the three letters C, M, and U.

First note that there may be a large, possibly even infinite number of geodesics connecting these three shapes of which we just depict the most intuitive ones. In principle, however, there are endless possibilities, for example, to split and remerge the letters in different ways and thereby achieve a locally shortest path. Which path is found by the algorithm depends on the initialisation of the intermediate shapes and deformations and

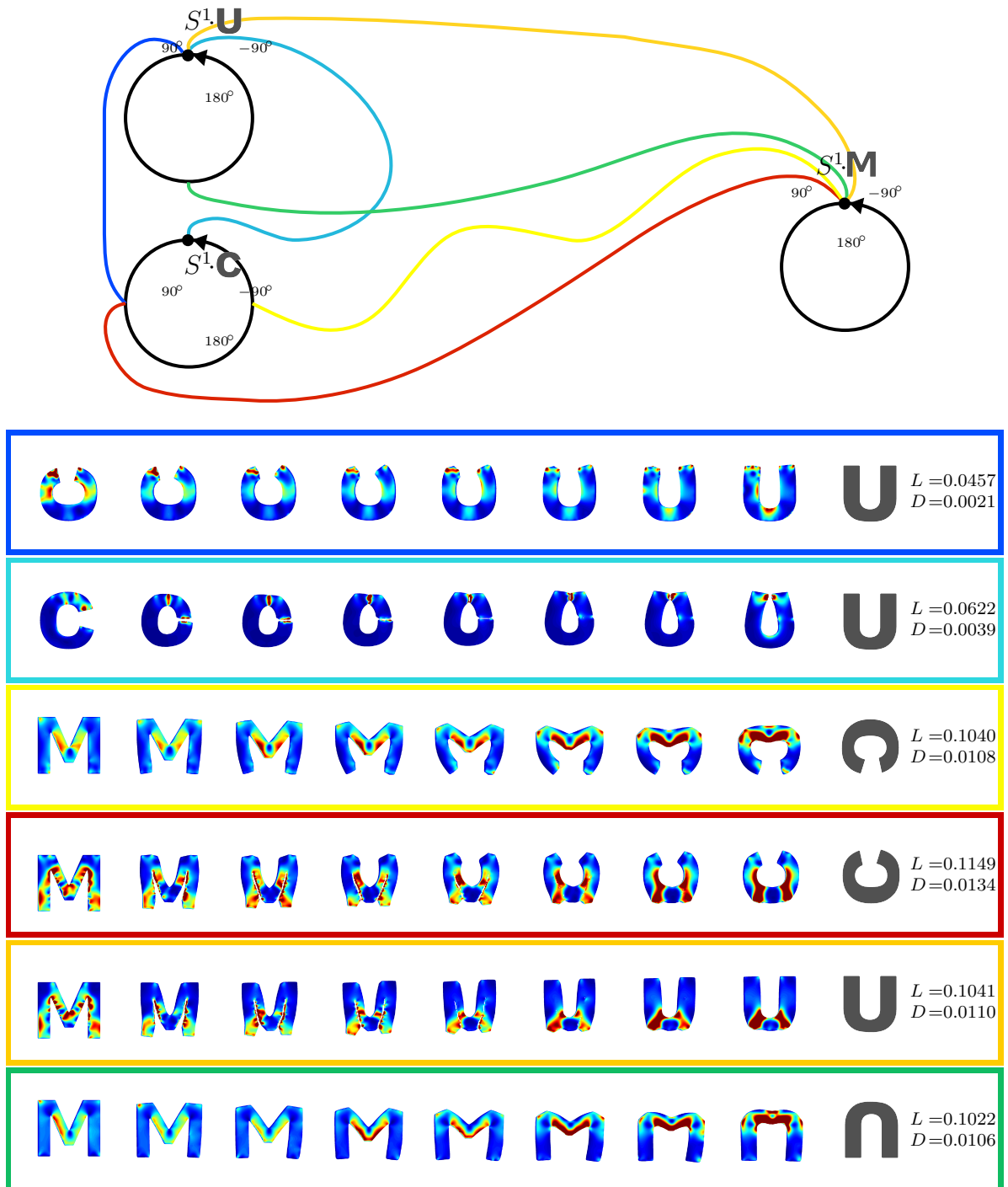


Figure 6.10: Sketch of shape space. The box around each geodesic and the corresponding path in the sketch are coloured accordingly. L denotes the geodesic length and D the total dissipation. The circles represent the action of S^1 on the shapes C , M , and U (that is, a rotation of the shapes), which induces no dissipation and has zero length.

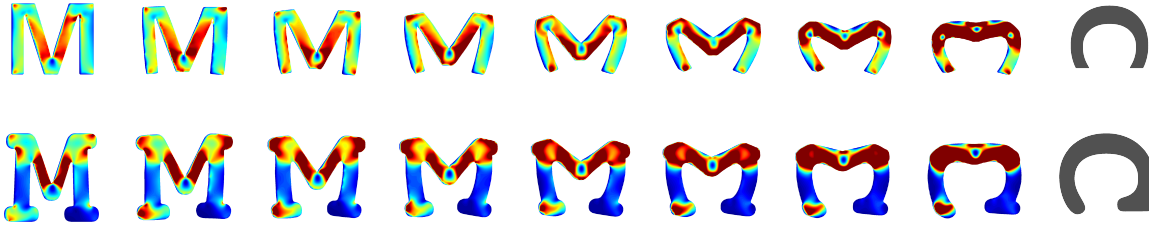


Figure 6.11: Discrete geodesics between an M and a C of a different font. The geodesic length (and total dissipation) are 0.1220 (0.01518) for the top row and 0.1276 (0.01634) for the bottom row.

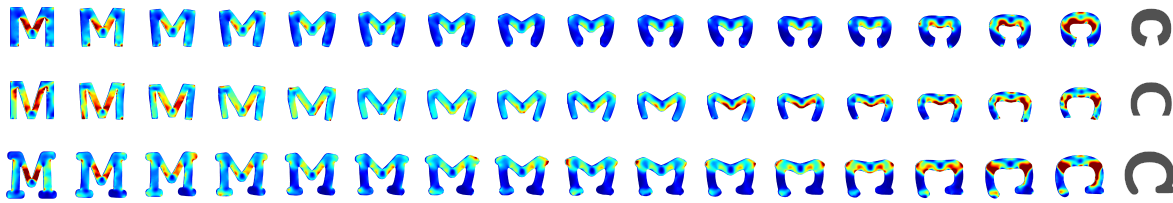


Figure 6.12: Finer time resolution of the geodesics between M and C from Figures 6.10 and 6.11. From top to bottom, the geodesic lengths (and the total dissipation) are 0.1025 (0.01056), 0.1201 (0.01465), and 0.1259 (0.01596).

therefore also on the position of the end shapes. This position is indicated in the sketch by the circle associated to each letter, which shall represent all possible rotation angles.

It is actually quite intuitive that the shortest geodesic between the C and the U involves a rotation by $\frac{\pi}{2}$ (top geodesic in Figure 6.10). Note that this rotation generates no dissipation so that the rotated and the upright C are identified with each other as being exactly the same shape; only the algorithm needs to start from the rotated C to detect this geodesic. Similarly, the shortest geodesics between M and C as well as between M and U are such that the inner two line segments of the M are bent outwards to yield a rotated C and U (third and last geodesic). If on the other hand there is crack formation or closure involved, the paths typically exhibit stronger dissipation near the cracks and are thus longer.

The robustness of this type of geodesics becomes apparent if we slightly perturb the end shapes. Of course, we then expect a similar geodesic path with similar intermediate shapes, a similar distribution of the dissipation, and a similar geodesic length. Figure 6.11 illustrates this continuous dependence of geodesics on the end shapes, where two more versions of the geodesic between M and C are computed.

As elaborated earlier, our rigid body motion invariant time discretisation is particularly intended to allow good approximations to a continuous geodesic already for relatively coarse time discretisations. Hence, as we refine a discrete geodesic by increas-

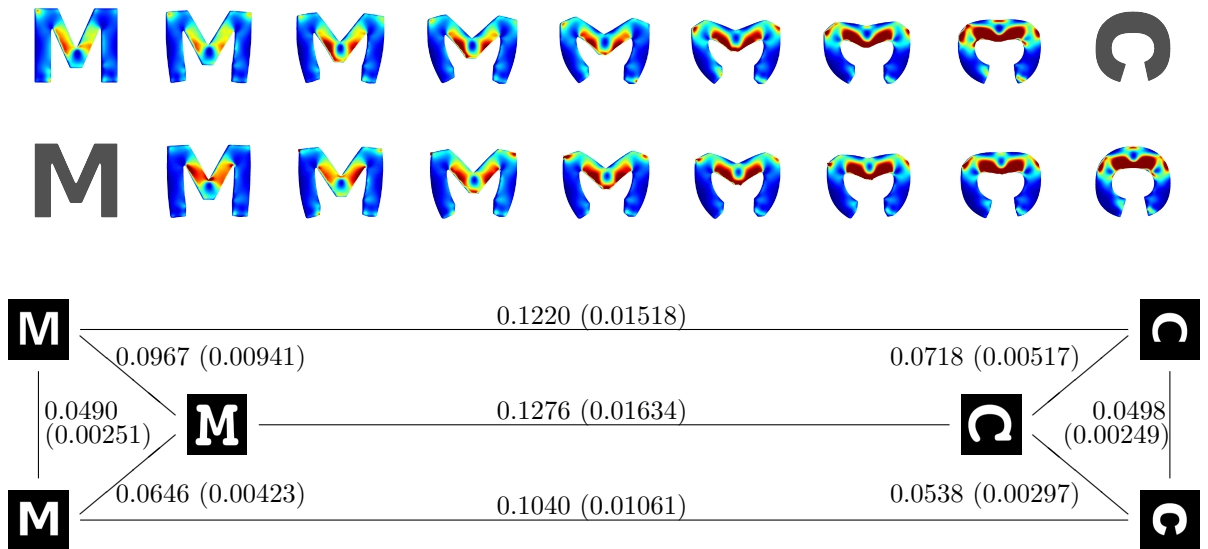


Figure 6.13: Top: Comparison of the discrete geodesic from M to C with the one from C to M . The geodesic lengths (total dissipation) are 0.1040 (0.01084) and 0.1030 (0.01064), respectively. Bottom: The geodesic lengths between the three M - and C -shapes satisfy the triangle inequality (values in brackets are total dissipation).

ing the number of time steps, we hope that the intermediate shapes do not change too strongly and that the geodesic length has already almost converged. Indeed, the geodesic length decreases by less than two per cent when halving the time step size for the different geodesics between M s and C s (Figure 6.12).

For sure, we also expect from a discrete geodesic to approximately satisfy the axioms of a metric, that is, that the geodesic distance between two shapes is roughly symmetric and satisfies the triangle inequality. These two properties are exemplarily illustrated in Figure 6.13.

6.5.2 Influence of material parameters

The model for relaxed discrete geodesics provides the flexibility to restrict the shape space to shapes of constant volume and finite perimeter (see Figure 6.6 for the impact of both terms). Furthermore, there is some flexibility associated with the chosen material parameters of viscous flow. If the ratio between the parameters λ and μ , which penalise local volume and length changes, respectively, is very small, then shapes deform along the geodesic by locally changing their volume as can be observed in Figure 6.14, top, where the dumb bell ends expand or shrink rather independently with almost no material flow between them. If volume changes are penalised strongly, however, the shape changes are achieved by a redistribution of the underlying material (bottom row).

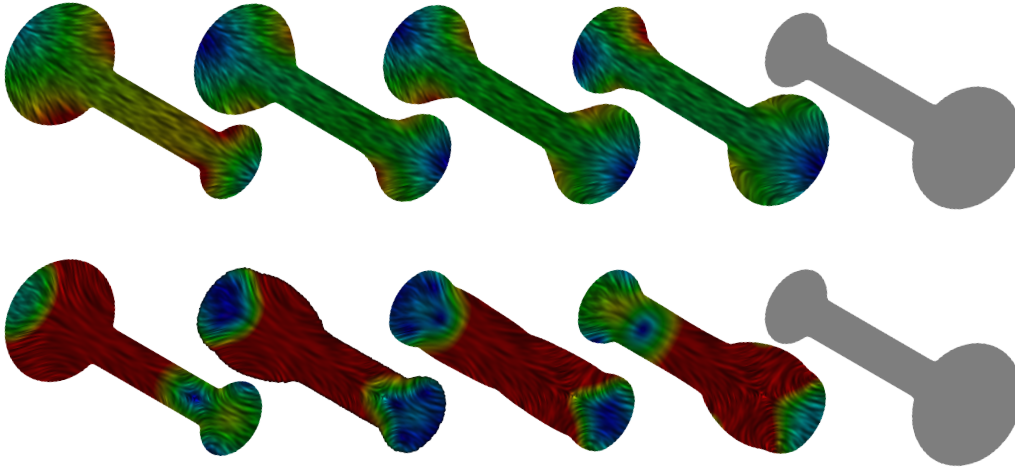



Figure 6.14: Two geodesic paths between dumb bell shapes varying in the size of the ends. In the first row, the ratio λ/μ between the parameters of the dissipation is 0.01 (leading to rather independent compression and expansion of the ends since the associated change of volume implies relatively low dissipation), and 100 in the second row (now mass is actually transported from one end to the other). The underlying texture on the shape domains $\mathcal{O}_0, \dots, \mathcal{O}_{K-1}$ is aligned to the transport direction, and the absolute value of the velocity field is colour-coded as .

6.5.3 Geodesics between partially occluded shapes

In some applications it might be desirable to compute a geodesic or a geodesic distance also between a partially occluded shape and, for example, a template shape. This can be accomplished by a small modification of the mismatch penalty \mathcal{F} . In Figure 6.15 we used the mismatch penalty

$$\tilde{\mathcal{F}}[\mathcal{O}_0, \phi_1, \mathcal{O}_1] = |\mathcal{O}_0 \setminus \phi_1^{-1}(\mathcal{O}_1)|$$

between the partially occluded and the deformed first shape. Here, we ensure that each point inside \mathcal{O}_0 has a counterpart in \mathcal{O}_1 , but not vice versa, so that the leg and the arm of the silhouette can be restored.

As level set approximation of the above mismatch penalty we employed

$$\tilde{\mathcal{F}}[u_0, \phi_1, u_1] = \int_{\Omega} H_{\varepsilon}(u_0) (H_{\varepsilon}(u_1(\phi_1)) - H_{\varepsilon}(u_0))^2 dx.$$

6.5.4 Geodesics between multilabelled images

The view of shapes solely as the outer boundaries $\mathcal{S} = \partial\mathcal{O}$ of open objects $\mathcal{O} \subset \mathbb{R}^d$ is rather limiting in some applications. While the contours of an object \mathcal{O}_A are correctly

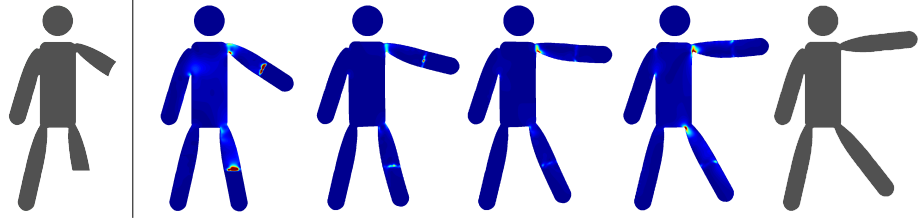


Figure 6.15: A discrete geodesic connecting different poses of a matchstick man can be computed (right of the vertical line), even though parts of one arm and one leg of the leftmost shape are occluded.

mapped onto the contours of an object \mathcal{O}_B via the geodesic between $\mathcal{S}_A = \partial\mathcal{O}_A$ and $\mathcal{S}_B = \partial\mathcal{O}_B$, the viscous fluid model imposes no restriction on the path-generating flow in the object interior (apart from the property that it should minimise the viscous dissipation). However, one might often want certain regions of one object \mathcal{O}_A to be mapped onto particular regions in another object \mathcal{O}_B . Generally speaking, realistic shapes or objects are often characterised as a composition of different structures or components that lie in a certain relative position to each other. A geodesic or a general path between two such shapes should of course match corresponding structures with each other, and a change in relative position of these subcomponents naturally has to contribute to the path length.

As an example, let us consider the discrete geodesic between the straight and the folded bar in Figure 6.4. The initial and the final shape contain no additional information about any internal structures so that the deformation strength and the induced dissipation along the geodesic path are distributed evenly over the whole material, producing symmetric intermediate shapes. However, if we prescribe the original and the final location for some internal region of the bar, the dissipation-minimising flow may look very different (if the additional constraints are not consistent with the geodesic flow without constraints, compare Figures 6.4 and 6.16).

For these reasons we would like to extend our approach to allow for more general shapes that may be composed of a number of subcomponents. Since we can interpret also images as collections of different shapes or objects, the computation of geodesics between (multilabelled) images nicely fits into this setting as well.

The extension is straightforward: Instead of a geodesic between just two shapes $\mathcal{S}_A = \partial\mathcal{O}_A$ and $\mathcal{S}_B = \partial\mathcal{O}_B$ or the corresponding objects, we now seek a geodesic path $(\mathcal{S}^i(t))_{i=1,\dots,n} = (\partial\mathcal{O}^i(t))_{i=1,\dots,n}$, $t \in [0, 1]$, between two collections of shapes, $(\mathcal{S}_A^i)_{i=1,\dots,n} = (\partial\mathcal{O}_A^i)_{i=1,\dots,n}$ and $(\mathcal{S}_B^i)_{i=1,\dots,n} = (\partial\mathcal{O}_B^i)_{i=1,\dots,n}$, generated by a joint motion field $v(t) : \bigcup_{i=1}^n \mathcal{O}^i(t) \rightarrow \mathbb{R}^d$. The single objects $\mathcal{O}^i(t)$ can then be regarded as the subcomponents of a large overall object $\bigcup_{i=1,\dots,n} \mathcal{O}^i(t)$. The total dissipation along the

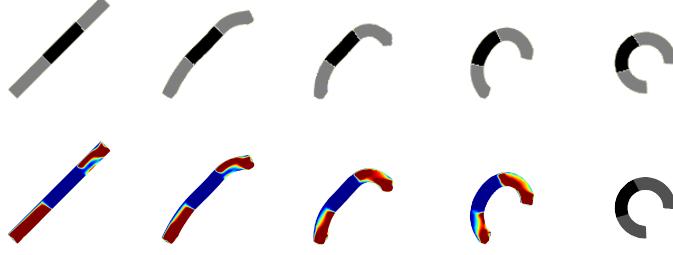


Figure 6.16: Discrete geodesic between the straight and the folded bar from Figure 6.4, where the white region of the initial shape in the top row is prescribed to be matched to the white region of the final shape. The bottom row shows a colour-coding of the corresponding viscous dissipation. Due to the strong difference in relative position of the white region between initial and end shape, the intermediate shapes exhibit a strong asymmetry and high dissipation at the bar ends.

path is measured exactly as before by

$$\mathbf{Diss}[v] = \int_0^1 \int_{\bigcup_{i=1}^n \mathcal{O}^i(t)} \frac{\lambda}{2} (\operatorname{tr} \epsilon[v])^2 + \mu \operatorname{tr}(\epsilon[v]^2) \, dx \, dt.$$

This naturally translates to the objective functional of the discrete geodesic with $K + 1$ intermediate shape collections $(\mathcal{S}_k^i)_{i=1, \dots, n}$, $k = 0, \dots, K$,

$$\sum_{k=1}^K \mathcal{W} \left[\phi_k, \bigcup_{i=1}^n \mathcal{O}_{k-1}^i \right] = \sum_{k=1}^K \int_{\bigcup_{i=1}^n \mathcal{O}_{k-1}^i} W(\mathcal{D}\phi_k) \, dx,$$

where the deformations $\phi_k : \bigcup_{i=1}^n \mathcal{O}_{k-1}^i \rightarrow \mathbb{R}^d$ satisfy the constraints $\phi_k(\mathcal{S}_{k-1}^i) = \mathcal{S}_k^i$ for $k = 1, \dots, K$, $i = 1, \dots, n$, and $\mathcal{S}_0^i = \mathcal{S}_A^i$, $\mathcal{S}_K^i = \mathcal{S}_B^i$, $i = 1, \dots, n$.

The corresponding relaxed formulation then has to employ multiple mismatch penalties (one for every constraint), and as before, we need a mild regularization of the shape perimeters so that the total energy of relaxed discrete geodesics between shape collections reads

$$\begin{aligned} & \mathcal{E}_\tau[(\phi_k, (\mathcal{S}_{k-1}^i)_{i=1, \dots, n}, (\mathcal{S}_k^i)_{i=1, \dots, n})_{k=1, \dots, K}] \\ &= \sum_{i=1}^K \left(\frac{1}{\tau} \mathcal{W} \left[\phi_k, \bigcup_{i=1}^n \mathcal{O}_{k-1}^i \right] + \sum_{i=1}^n (\gamma \mathcal{F}[\mathcal{O}_{k-1}^i, \phi_k, \mathcal{O}_k^i] + \eta \tau \mathcal{L}[\mathcal{S}_k^i] + \nu \tau \mathcal{V}[\mathcal{O}_k^i]) \right). \end{aligned} \quad (6.1)$$

For sure, the different object components \mathcal{O}_A^i or \mathcal{O}_B^i may overlap, but they have to do so consistently in the initial collection of shapes and the final one, that is, there must exist a flow that deforms $(\mathcal{O}_A^i)_{i=1, \dots, n}$ into $(\mathcal{O}_B^i)_{i=1, \dots, n}$. In fact, it is often desired



Figure 6.17: Top: Real frames from a video sequence. Middle: Discrete geodesic between the first and the last segmented frame with three and with seven intermediate steps. Bottom rows: Pullback of the last frame (top) and pushforward (bottom) of the first one (the background has been pasted into the pullbacks and pushforwards so that it is not deformed).

that the different objects overlap: Assume \mathcal{O}^1 and \mathcal{O}^2 to be disjoint but have a common boundary. Obviously, it costs zero energy to pull both objects apart rigidly. Hence, if \mathcal{O}^1 and \mathcal{O}^2 shall keep the common boundary along paths in shape space, one of the objects should be replaced by the interior of $\overline{\mathcal{O}^1 \cup \mathcal{O}^2}$ so that a separation of both components first requires the costly generation of a new boundary. For this reason we have composed the object in Figure 6.16 of two objects, one representing the whole bar and the other the black region. Another example is given in Figure 6.17, where the head and the torso served as one component and the torso and the legs as a second one.

Rephrasing the above energy in terms of level set functions is straightforward, and the approximations of the different energy terms have already been stated earlier. Note that with n level set functions and thus n object components \mathcal{O}^i we can in fact distinguish 2^n different phases which represent the pure objects \mathcal{O}^i , $i = 1, \dots, n$, as well as all possible combinations of overlappings. For example, four phases (head, torso, legs, background) have been described using two level set functions in Figure 6.17. Of course, it is furthermore possible to assign each phase different material properties. This has been pursued in Figure 6.18, where a geodesic between two frames from a video of blood cells has been computed. The top row shows frames from the real video sequence, where

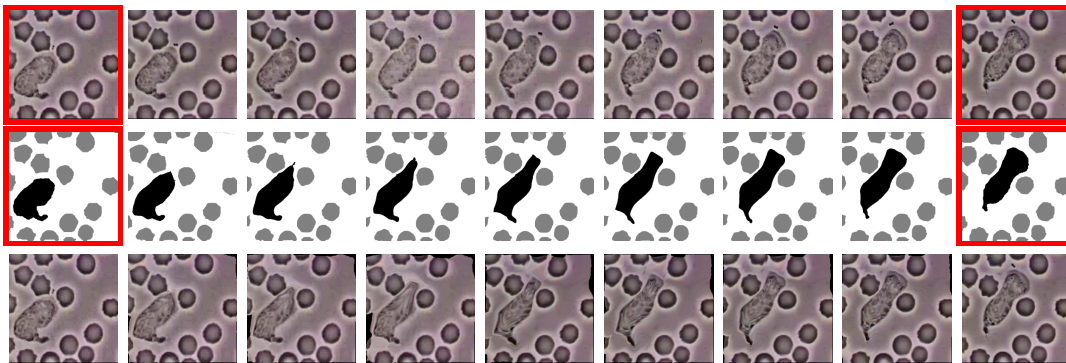


Figure 6.18: Top: Frames from a real video sequence of a white blood cell among a number of red ones (courtesy Robert A. Freitas, Institute for Molecular Manufacturing, California, USA). Middle: Computed discrete multiphase geodesic between the first and the last frame. Bottom: Pushforward of the initial (first four shapes) and pullback (last five shapes) of the final frame according to the geodesic flow.

a white blood cell squeezes through a number of red blood cells. For the computation of the geodesic (middle row), we employed two level set functions and assigned the white blood cell with material parameters twenty times weaker than for the red blood cells (material parameters of the background are 10^4 times weaker). This seems reasonable, given that red blood cells are comparatively stiff.

6.5.5 Shape clustering via geodesic distances

As a final application we consider shape clustering according to geodesic path length, the first example being a set of six 3D foot shapes (Figure 6.19, left). Although the shapes look very similar, one can clearly separate the first three feet from the rest. Finally, we have evaluated distances between different 2D letters, and the resulting clustering is depicted in Figure 6.19, right.

6.6 Discussion

So far, topological transitions do not influence the energy of the geodesic morphing path directly. They just result in additional costs since the associated deformations are usually strong ones. However, we have seen that the lack of a penalisation of topological changes might lead to phenomena like crack formation. In order to reduce such effects, one could introduce a cost for the change of perimeter length, which would correspond to a dissipation associated with the formation or disappearance of interfaces.

Another possibility to regularise the shapes along the path consists in the use of prior

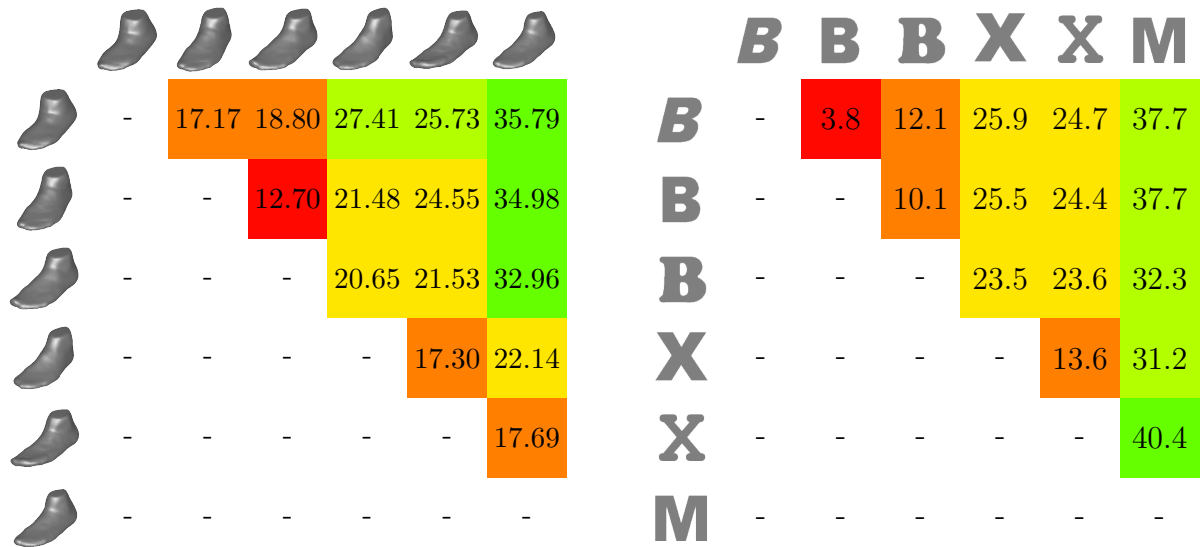


Figure 6.19: Left: Pairwise geodesic distances between different 3D foot shapes (data courtesy of adidas). A cluster of the first three feet is clearly visible. Right: Pairwise geodesic distances between (also topologically) different letter shapes. Obviously, the Bs and Xs form clusters, and these two clusters are closer to each other than the significantly distant M.

statistical knowledge. The further the intermediate shapes are away from, for example, some submanifold of learnt shapes, the more expensive the path becomes. The viscous Riemannian metric would then be changed to a metric that yields shorter paths within or close to this submanifold than outside.

Concerning the numerical relaxation it might be of interest to implement a multigrid-type scheme in space and time with V-cycles of energy relaxation from fine to coarse time and space resolutions and back. In this way, a fast flow of information can be ensured on the coarse level, while little details are treated at the corresponding fine level. If these local details, which did not appear on the coarser scale yet, influence the global path (for example, an interlocking between two shape structures), then the corresponding information can be passed on more quickly when returning to the coarse resolution.

Finally, possible next steps also include some implementation of time adaptivity: From the local change of geodesic path length, induced by locally refining the time discretisation, we may draw conclusions about the proximity to the time-continuous geodesic and use this information to adaptively refine the path.

7 Minimum compliance design in nonlinear elasticity

This chapter is concerned with the optimisation of an object or a domain \mathcal{O} with respect to its structural rigidity. More precisely, we would like to find $\mathcal{O} \subset \Omega$ (for some open connected set $\Omega \subset \mathbb{R}^d$) such that \mathcal{O} deforms as little as possible if subjected to a fixed, prescribed surface load. The strength of the deformation is either measured as the change in potential energy of the surface load or as the total elastic deformation energy accommodated by \mathcal{O} (or equivalently, the work performed by the surface load), both of which describe a different type of the so-called compliance.

Minimum compliance design has applications in engineering where mechanical devices or structural components have to be developed that optimally balance material consumption or component weight with the component stiffness or rigidity. Since the compliance can also be interpreted as the mechanical energy that has to be absorbed by the structure, minimising the compliance is related to reducing the risk of material failure.

We will propose a phase field model to design shapes which minimise a weighted sum of their volume, their perimeter, and their compliance. Since the distinction between material and void is central to the determination of the mechanical shape properties, we will choose a double well phase field of Modica–Mortola type (compare Section 3.2). Furthermore, instead of performing the shape optimisation in the setting of linearised elasticity (as is typically done in the literature, see Section 2.4), we will employ a nonlinear hyperelastic constitutive law to describe the shape deformations and the associated compliance. We will also analyse the existence of optimal, shape-encoding phase fields as well as the model behaviour for decreasing phase field interface width.

The behaviour of the optimisation system with nonlinear elasticity differs significantly from shape optimisation in linearised elasticity. To begin with, there exist several possible definitions for the compliance which are all equivalent in linearised elasticity but now result in very different optimisation problems. We may consider the change in potential energy of the surface load, the stored elastic deformation energy, or the dissipation associated with the deformation. Furthermore, the symmetry property is lost that a sign change of the surface loads has no influence on the optimal shape. As a consequence, shape optimisation problems that yield symmetric shapes in linearised elasticity now result in asymmetric shapes. Additionally, the deformation induced by the surface load is generally no longer unique, which complicates the mathematical analysis. Buckling instabilities may appear in which structures such as compressed beams can

bend to either side, thereby producing nonuniqueness. Since different buckling deformations generally correspond to different compliance values, one may experience that the compliance suddenly jumps up during shape optimisation as the global equilibrium deformation switches from one buckling deformation to another one. This phenomenon may result in nonexistence of minimisers if we always pick the worst case from the set of all equilibrium deformations. Finally, also numerically, the use of nonlinear elasticity poses a challenge. We typically observe rather large, geometrically strongly nonlinear deformations. Their computation requires robust numerical minimisation methods that reliably detect local rotations and bypass saddle points which frequently appear between two buckling deformations.

After presenting the optimisation problem and its phase field approximation in Section 7.1 we will briefly examine the peculiarities associated with the use of nonlinear elasticity in Section 7.2. The existence of minimisers and the sharp interface limit of the phase field model are studied in Section 7.3 before stating the implementation in Section 7.4 and finally showing few experiments in Section 7.5.

7.1 Geometry optimisation for a prescribed mechanical load

Let us briefly recapitulate the mechanical framework from Section 3.1. Assume we are given a sufficiently regular, elastic body $\mathcal{O} \subset \mathbb{R}^d$ which is fixed at part of its boundary, $\Gamma_D \subset \partial\mathcal{O}$, and subjected to a sufficiently regular surface load $F : \Gamma_N \rightarrow \mathbb{R}^d$ on $\Gamma_N \subset \partial\mathcal{O}$ (compare Figure 3.1). The body obviously deforms under the surface load, and the equilibrium deformation $\phi : \mathcal{O} \rightarrow \mathbb{R}^d$ minimises the total free energy

$$\mathcal{E}[\mathcal{O}, \phi] = \mathcal{W}[\mathcal{O}, \phi] - \mathcal{C}[\phi]$$

among all deformations $\phi \in W^{1,p}(\mathcal{O})$ with trace $\phi|_{\Gamma_D} = \text{id}$ (if W satisfies a growth condition of order p), where

$$\mathcal{W}[\mathcal{O}, \phi] = \int_{\mathcal{O}} W(\mathcal{D}\phi) \, dx$$

describes the elastic energy stored inside the material and

$$\mathcal{C}[\phi] = \int_{\Gamma_N} F \cdot (\phi - \text{id}) \, da$$

is the negative potential of the surface load. In our computations, we will employ the particular material law $W(\mathcal{D}\phi) = \frac{\mu}{2} \|\mathcal{D}\phi\|_F^2 + \frac{\lambda}{4} \det \mathcal{D}\phi^2 - (\mu + \frac{\lambda}{2}) \log \det \mathcal{D}\phi - \frac{d\mu}{2} - \frac{\lambda}{4}$ for material parameters λ and μ (compare Section 3.1).

In linearised elasticity, where the energy density of the material is a quadratic function $W(\mathcal{D}\phi) = \frac{1}{2} \mathbf{C}(\mathcal{D}\phi - I) : (\mathcal{D}\phi - I)$ of the strain or deformation gradient $\mathcal{D}\phi$ (for some

symmetric positive semi-definite elasticity tensor \mathbf{C}), we obtain the equilibrium condition $0 = \int_{\mathcal{O}} \mathbf{C}(\mathcal{D}\phi - I) : \mathcal{D}\theta \, dx - \int_{\Gamma_N} F \cdot \theta \, da$ for all test displacements θ . In particular, this holds for $\theta = \phi - \text{id}$, which implies $2\mathcal{W}[\mathcal{O}, \phi] = \mathcal{C}[\phi]$ for the equilibrium deformation ϕ . Here, $2\mathcal{W}[\mathcal{O}, \phi] = \mathcal{C}[\phi]$ represents a measure of the deformation strength and is denoted the compliance of the object \mathcal{O} , which may be seen as some kind of inverse rigidity. For a maximally rigid structure with least energy absorption, this compliance is usually minimised under some additional volume and regularity constraints (see Section 2.4).

During our shape optimisation, we would also like to find $\mathcal{O} \subset \Omega \subset \mathbb{R}^d$ such that its compliance is minimised for a given surface load F . In linearised elasticity, it does obviously not matter whether we describe the compliance via $2\mathcal{W}[\mathcal{O}, \phi]$ or $\mathcal{C}[\phi]$. However, we will use nonlinear hyperelastic constitutive laws so that—as we will see later—it makes a difference which term we choose to minimise. In the following, we will hence state the model for both choices.

7.1.1 Balancing compliance with volume and perimeter

If we look for $\mathcal{O} \subset \Omega \subset \mathbb{R}^d$ such that just the compliance $\mathcal{W}[\mathcal{O}, \phi]$ or $\mathcal{C}[\phi]$ is minimised, then, trivially, $\mathcal{O} \equiv \Omega$ certainly yields the most rigid structure. Hence, we are actually interested in a balance between rigidity and material consumption (or weight), expressed as the Lebesgue measure of \mathcal{O} ,

$$\mathcal{V}[\mathcal{O}] = |\mathcal{O}|.$$

However, domains \mathcal{O} generally do not exist that minimise a weighted sum of compliance and volume. Typically, microstructures form along a minimising sequence \mathcal{O}_i , $i \in \mathbb{N}$, in particular rank- d sequential laminates in which material and void rapidly alternate (see Section 2.4). This behaviour is associated with a weak, but not strong convergence of the characteristic functions $\chi_{\mathcal{O}_i}$. The optimisation problem can be turned into a well-posed one by regularisation, for which there are several possibilities (presented in Section 2.4). We choose to replace the void, $\Omega \setminus \mathcal{O}$, by a weak material with a stiffness reduced by a factor $\delta \ll 1$ and to add the domain perimeter,

$$\mathcal{L}[\partial\mathcal{O}] = \mathcal{H}^{d-1}(\partial\mathcal{O}),$$

as a regularising prior, which can be interpreted as introducing manufacturing costs for the production and processing of the object surface. The substitution of void by a weak material is achieved by replacing $\mathcal{W}[\mathcal{O}, \phi]$ by

$$\mathcal{W}^\delta[\mathcal{O}, \phi] = \int_{\Omega} ((1 - \delta)\chi_{\mathcal{O}} + \delta)W(\mathcal{D}\phi) \, dx$$

(for the characteristic function $\chi_{\mathcal{O}}$ of \mathcal{O}) and the total free energy $\mathcal{E}[\mathcal{O}, \phi]$ by

$$\mathcal{E}^\delta[\mathcal{O}, \phi] = \mathcal{W}^\delta[\mathcal{O}, \phi] - \mathcal{C}[\phi],$$

where we choose $\delta = 10^{-4}$ in our computations. As in the previous chapters, this allows to properly define the deformation ϕ also outside \mathcal{O} and (combined with suitable Dirichlet boundary conditions on $\partial\Omega$) to prevent self-interpenetration of matter in form of overlapping material parts. Moreover, this procedure has been justified at least in the case of shape optimisation via the homogenisation method [3]. The final optimisation problem that we will consider then is to minimise (for parameters $\eta, \nu > 0$) either

$$\begin{aligned} \mathcal{J}_W[\mathcal{O}, \phi] &= 2\mathcal{W}^\delta[\mathcal{O}, \phi] + \nu\mathcal{V}[\mathcal{O}] + \eta\mathcal{L}[\partial\mathcal{O}] \\ \text{or } \mathcal{J}_C[\mathcal{O}, \phi] &= \mathcal{C}[\phi] + \nu\mathcal{V}[\mathcal{O}] + \eta\mathcal{L}[\partial\mathcal{O}] \end{aligned}$$

for $\mathcal{O} \in \{\mathcal{O} \subset \Omega : \Gamma_D, \Gamma_N \subset \partial\mathcal{O}\}$ under the constraint that $\phi : \Omega \rightarrow \mathbb{R}^d$ minimises the free energy $\mathcal{E}^\delta[\mathcal{O}, \phi]$ among all deformations in $W^{1,p}(\Omega)$ whose trace is the identity on Γ_D .

7.1.2 Allen–Cahn phase field approximation

We have already mentioned in various places that the optimisation of an explicitly represented object \mathcal{O} is numerically difficult. For this reason we will approximate \mathcal{O} with a double well phase field $u : \Omega \rightarrow \mathbb{R}$ of Modica–Mortola- or Allen–Cahn-type. Such phase fields constitute a convenient implicit representation of objects and allow for a simple approximation of their boundary length. They originate from the description of biphasic materials in physics: Each phase corresponds to a value of u , for example, $u = -1$ might indicate one and $u = 1$ the other phase (instead of -1 and 1 one might just as well choose the concentration of a chemical constituent in each phase and interpret u at any point $x \in \Omega$ as the local concentration of that constituent). Intermediate values signify impure regions, for example, at the phase interfaces. The local chemical bulk energy density is a function $\Psi(u)$ of u , whose global minima we may assume to be $\Psi(-1) = \Psi(1) = 0$, so that the total bulk energy reads $\int_\Omega \Psi(u) dx$. This energy is perturbed by an interfacial energy of the form $\int_\Omega |\nabla u|^2 dx$. By a rescaling, we obtain the free energy

$$\mathcal{L}_{\text{MM}}^\varepsilon[u] = \frac{1}{2} \int_\Omega \varepsilon |\nabla u|^2 + \frac{1}{\varepsilon} \Psi(u) dx,$$

which for $\varepsilon \rightarrow 0$ forces the phase field u towards the pure phases -1 and 1 and Γ -converges against a multiple of the total interface length (see Section 3.2). For finite ε , the interface is not sharp, but represented by a smooth transition layer, whose width scales with ε . We will distinguish the two pure phases void and material.

In the following, we shall assume the phase $u = 1$ to represent the inside of \mathcal{O} and $u = -1$ the outside. An intermediate value will indicate a phase interface, and values outside $[-1, 1]$ are not allowed. The perimeter term $\mathcal{L}[\partial\mathcal{O}]$ is then replaced by $\mathcal{L}_{\text{MM}}^\varepsilon[u]$, where we choose the double well potential

$$\Psi(u) = \frac{9}{16}(u^2 - 1)^2.$$

Furthermore, we introduce an approximation $\chi_{\mathcal{O}}[u]$ to the characteristic function $\chi_{\mathcal{O}}$, for example,

$$\chi_{\mathcal{O}}[u] = \frac{1}{4}(u + 1)^2.$$

With this characteristic function at hand, we approximate the total volume $\mathcal{V}[\mathcal{O}]$ as

$$\mathcal{V}[u] = \int_{\Omega} \chi_{\mathcal{O}}[u] \, dx$$

and the stored elastic energy as

$$\mathcal{W}[u, \phi] = \int_{\Omega} ((1 - \delta)\chi_{\mathcal{O}}[u] + \delta)W(\mathcal{D}\phi) \, dx$$

(where we dropped the superscript δ for simplicity). Overall, we will minimise the functional

$$\begin{aligned} \mathcal{J}_{\mathcal{W}}^{\varepsilon}[u, \phi] &= 2\mathcal{W}[u, \phi] + \nu\mathcal{V}[u] + \eta\mathcal{L}_{\text{MM}}^{\varepsilon}[u] \\ \text{or } \mathcal{J}_{\mathcal{C}}^{\varepsilon}[u, \phi] &= \mathcal{C}[\phi] + \nu\mathcal{V}[u] + \eta\mathcal{L}_{\text{MM}}^{\varepsilon}[u] \end{aligned}$$

for integrable functions $u : \Omega \rightarrow [-1, 1]$ with $u|_{\Gamma_D \cup \Gamma_N} = 1$ under the constraint that $\phi : \Omega \rightarrow \mathbb{R}^d$ with $\phi|_{\Gamma_D} = \text{id}$ minimises the energy

$$\mathcal{E}[u, \phi] = \mathcal{W}[u, \phi] - \mathcal{C}[\phi].$$

7.2 Effect of nonlinear elasticity in shape optimisation

The use of nonlinear instead of linearised elasticity changes the nature of the compliance minimisation problem qualitatively. In the following, we will briefly discuss the influence of different compliance definitions (which are equivalent only in the case of linearised elasticity) on the shape optimisation, the symmetry breaking effect of nonlinear elasticity on the optimal shapes, and the problems associated with the existence of buckling instabilities.

7.2.1 Choice of compliance definition

As already explained earlier, the compliance of an object \mathcal{O} may be regarded as a kind of inverse rigidity and can in linearised elasticity be described as $2\mathcal{W}[\mathcal{O}, \phi]$ or equivalently $\mathcal{C}[\phi]$ for the equilibrium deformation ϕ . In nonlinear elasticity, however, $\mathcal{W}[\mathcal{O}, \phi]$ and $\mathcal{C}[\phi]$ are no longer related by a factor of 2, and the question arises which one appropriately corresponds to the compliance in the linearised setting and which one should be chosen

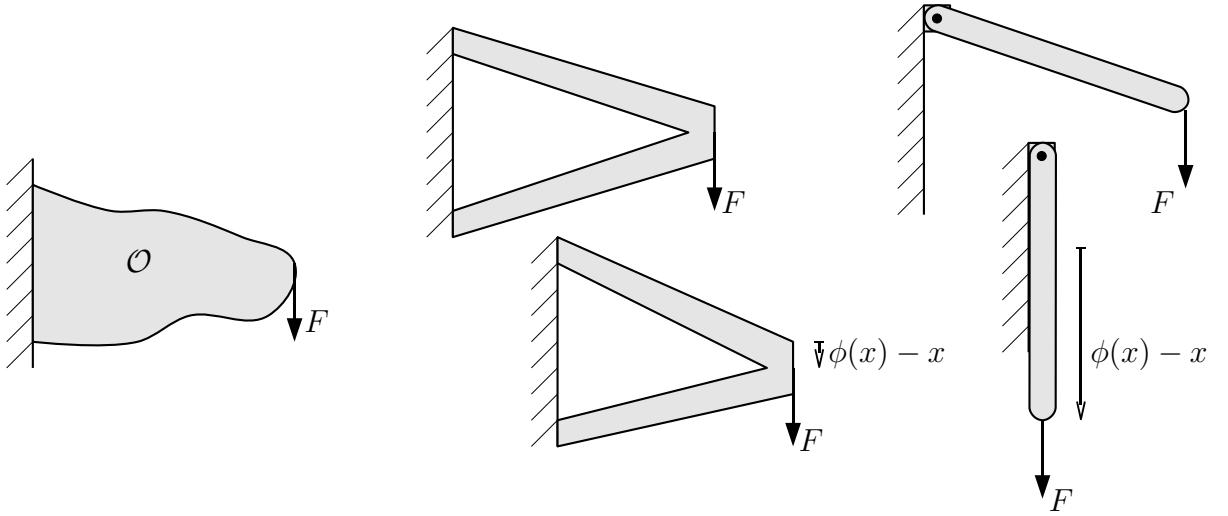


Figure 7.1: Left: Sketch of the design problem. We aim to find an optimal structure \mathcal{O} to bear the load F . Middle, right: Two possible designs. The middle design exhibits rather low $\mathcal{C}[\phi]$, but high $\mathcal{W}[\mathcal{O}, \phi]$ whereas the right design (with a hinge) yields the reverse.

for shape optimisation. The stored elastic energy $\mathcal{W}[\mathcal{O}, \phi]$ corresponds to the work transferred to the body \mathcal{O} by the surface load, while the total decrease $\mathcal{C}[\phi]$ in the potential energy of the surface load F is composed of exactly this work plus the energy dissipation during the system dynamics before the equilibrium is reached. Allaire et al., who have already computed one nonlinearly elastic shape optimisation example [6], employ the surface load potential and try to minimise $\mathcal{C}[\phi]$. In fact, the other choice is possible, too, and both yield qualitatively different results.

A simple model example shall illustrate the conceptual differences: Consider a design task as in Figure 7.1, left, where a structure \mathcal{O} has to bear a load F . A cantilever-like design (middle sketch) exhibits a rather small displacement $\phi - \text{id}$ and thus a small value of $\mathcal{C}[\phi]$, but the strong compression of the lower branch causes a relatively high deformation energy $\mathcal{W}[\mathcal{O}, \phi]$. A freely rotating rod, on the other hand, allows a strong displacement with high $\mathcal{C}[\phi]$ but low $\mathcal{W}[\mathcal{O}, \phi]$. The former design is more appropriate if the load is supposed to be sustained without large displacements while the latter design is more related to the material strain and is useful in systems where the energy dissipation on the way to the final equilibrium configuration is absorbed in a reasonable way (if, for example, the structure is embedded in a viscous fluid). We will consider both choices, but we have to keep in mind that shape optimisation with respect to $\mathcal{C}[\phi]$ will yield results of the same type as in Figure 7.1, middle, while optimisation with respect to $\mathcal{W}[\mathcal{O}, \phi]$ generally allows strong deformations and tends to produce shapes as in Figure 7.1, right.

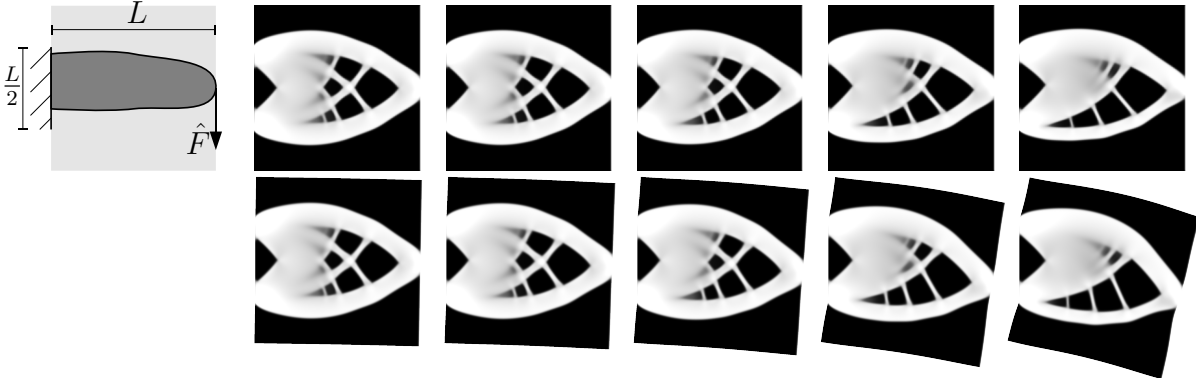


Figure 7.2: Optimal design of a cantilever according to the sketch top left. The top row shows optimal designs for loads $\hat{F} = 0.5, 1, 2, 4, 6$ (the point load \hat{F} is approximated by a tent-like surface load F along a width of 2^{-3}) and $\eta = 2^5 \cdot 10^{-5} \cdot \hat{F}^2$, $\nu = 2^{10} \cdot 10^{-4} \cdot \hat{F}^2$ (resolution 257^2 , $\lambda = \mu = 80$, $L = 1$). In linearised elasticity, all parameter combinations would yield exactly the same symmetric, optimal shape whereas here, we see asymmetries evolving. Computations were performed using the phase field model. White indicates full material while black represents a 10^4 times weaker material.

7.2.2 Proper handling of load asymmetries

Another feature that distinguishes nonlinear from linearised elasticity is that the sign of load F has a nonlinear impact on the deformation and thus also on the optimal geometry \mathcal{O} . In linearised elasticity, the (unique) equilibrium deformation is the minimiser of the free energy

$$\mathcal{E}_F^{\text{lin}}[\mathcal{O}, \phi] = \mathcal{W}^{\text{lin}}[\mathcal{O}, \phi] - \mathcal{C}_F[\phi] = \int_{\mathcal{O}} \frac{1}{2} \mathbf{C}(\mathcal{D}\phi - I) : (\mathcal{D}\phi - I) \, dx - \int_{\Gamma_N} F \cdot (\phi - \text{id}) \, da$$

for the symmetric, positive semi-definite elasticity tensor \mathbf{C} , where the subscript F indicates the surface load. Obviously, if ϕ_F minimises $\mathcal{E}_F^{\text{lin}}[\mathcal{O}, \cdot]$, then $\phi_{-F} := 2\text{id} - \phi_F$ minimises $\mathcal{E}_{-F}^{\text{lin}}[\mathcal{O}, \cdot]$, the total free energy where the direction of the surface load has been reversed. Furthermore, $\mathcal{E}_F^{\text{lin}}[\mathcal{O}, \phi_F] = \mathcal{E}_{-F}^{\text{lin}}[\mathcal{O}, \phi_{-F}]$, $\mathcal{W}^{\text{lin}}[\mathcal{O}, \phi_F] = \mathcal{W}^{\text{lin}}[\mathcal{O}, \phi_{-F}]$, and $\mathcal{C}_F[\phi_F] = \mathcal{C}_{-F}[\phi_{-F}]$. As a consequence, the optimal geometry \mathcal{O} for a prescribed load F is the same one as for the load $-F$, that is, a sign change of the load has no influence on the shape optimisation. In addition, if the sign change of F has the same effect as mirroring the shape optimisation problem (for example, for the cantilever design in Figure 7.2), then the resulting optimal shapes are symmetric.

In contrast, in nonlinear elasticity, the material behaviour and geometry change strongly depend on whether we tear at or push against an object. A sign change of the load F no longer simply implies a sign change of the displacement. Consequently, the symmetry property of linearised elasticity is lost: Where the shape optimisation with

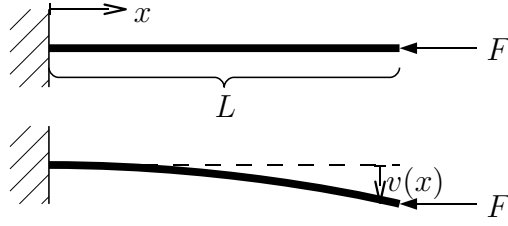


Figure 7.3: Undeformed and deformed (buckled) configuration of a compressed beam.

linearised elasticity results in a symmetric shape \mathcal{O} , the corresponding optimal geometry for nonlinear elasticity is generally asymmetric. This phenomenon can already be observed for quite small displacements as shown in Figure 7.2 for the design of a cantilever, where the effect of gradually increasing the load F is explored.

7.2.3 Buckling instabilities

There is yet a further striking phenomenon which is exhibited by nonlinearly elastic systems and cannot be captured by linearised elasticity. It is associated with the non-uniqueness of equilibrium deformations: While in linearised elasticity, the total free energy $\mathcal{E}[\mathcal{O}, \phi]$ is convex and quadratic in the deformation ϕ , the energy landscape is much more complicated in nonlinear elasticity and generally admits multiple (locally) minimising deformations ϕ . Of course, this raises the serious question which equilibrium deformations we should actually consider during shape optimisation, and we will gain some insight into this issue in Section 7.3.2.

Typically, the multiple locally minimising deformations involve local bending of structures, and the classical example is given by the compression of straight bars. Consider a straight horizontal bar which is clamped at its left end and subjected to a horizontal compression load F at its right end (Figure 7.3). Let us denote the vertical displacement at position $x \in [0, L]$ by $v(x)$, then the bending moment $M(x)$ inside the bar at x is given by

$$M(x) = F(v(L) - v(x)) .$$

Under the assumption of Bernoulli's beam hypothesis and Hooke's law with Young's modulus E , this moment can also be expressed as $M(x) = \frac{EI}{\rho}$, where I denotes the second moment of cross-sectional area and ρ is the radius of bending curvature. Upon approximating $\frac{1}{\rho} \approx \frac{\partial^2 v(x)}{\partial x^2}$ we obtain the linear ordinary differential equation

$$EI \frac{\partial^2 v(x)}{\partial x^2} = F(v(L) - v(x)) ,$$

which together with the boundary conditions $v(0) = 0$ and $\frac{\partial v(0)}{\partial x} = 0$ can be solved as

$$v(x) = v(L) \left(1 - \cos \left(\sqrt{\frac{F}{EI}} x \right) \right)$$

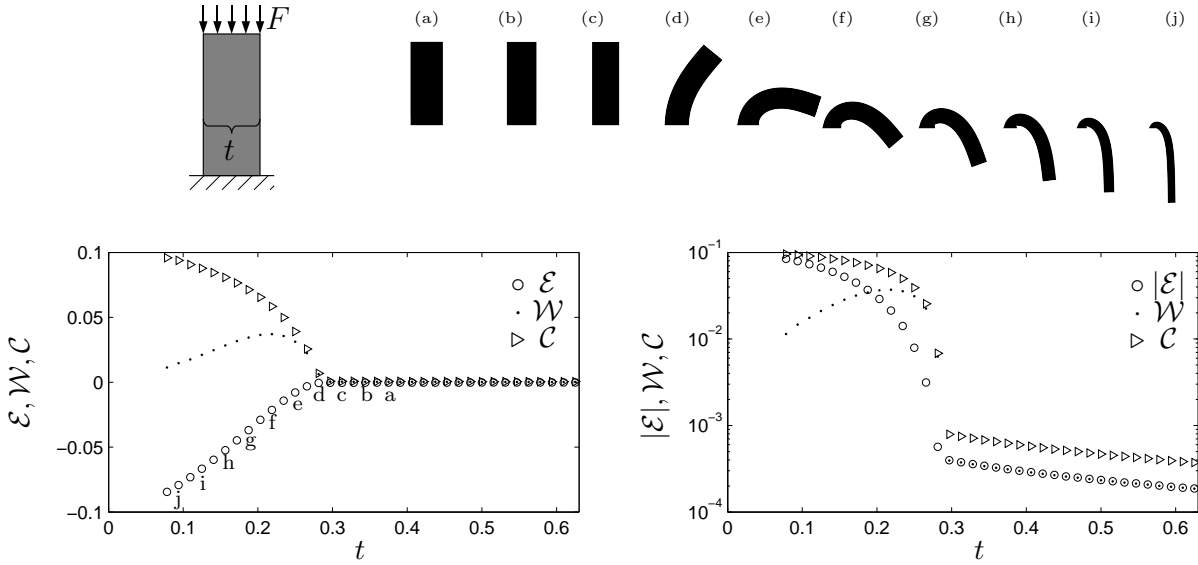


Figure 7.4: Top: Sketch of the load configuration (left) and corresponding equilibrium deformations for beams of varying width (from (a) to (j)). Bottom: $\mathcal{E}[\mathcal{O}, \phi]$, $\mathcal{W}[\mathcal{O}, \phi]$, and $\mathcal{C}[\phi]$ for the equilibrium deformations as a function of beam thickness t (the second graph shows a logarithmic plot).

Hence, $v(L)$ may be nonzero for $\sqrt{\frac{F}{EI}}L = \frac{\pi}{2} + n\pi$, $n \in \mathbb{Z}$, and for $n = 0$ we obtain the buckling load $F = \frac{EI\pi^2}{4L^2}$, that is, the smallest load for which we expect a bending of the beam towards one side rather than a symmetric compression. The same analysis for different boundary conditions yields the so-called Euler buckling modes.

The physical bifurcation associated with buckling of beams can be reproduced in a nonlinear elasticity model. Figure 7.4 shows simulation results for the compression of vertical bars with height 1 and varying thickness t . The top edge of each bar is subjected to a uniformly distributed surface load such that the total resulting downward force is the same for all bars. The mechanical energy components belonging to the different configurations are shown in Figure 7.4, right, as functions of the bar thickness t . Apparently, down to a width of $t = \hat{t} \approx 0.28$, we seem to stay in the linearly elastic regime: The deformations ϕ of the beams \mathcal{O} are symmetric, and $\mathcal{W}[\mathcal{O}, \phi] \approx \frac{1}{2}\mathcal{C}[\phi] \approx -\mathcal{E}[\mathcal{O}, \phi]$ as in linear elasticity. For smaller thicknesses t , all energy components strongly increase, and the beams bend outwards.

Note that there is a beam width \tilde{t} below which the stored elastic energy $\mathcal{W}[\mathcal{O}, \phi]$ decreases again. This behaviour is linked to the observation from the previous sections concerning the difference between $\mathcal{W}[\mathcal{O}, \phi]$ and $\mathcal{C}[\phi]$. The thinner the bottom of the beam the less bending energy is stored, and its base behaves more like a hinge so that the entire configuration resembles just a hanging, dilated rod, which absorbs relatively little elastic energy.

For the example of a compressed beam, the symmetric state, which corresponds to

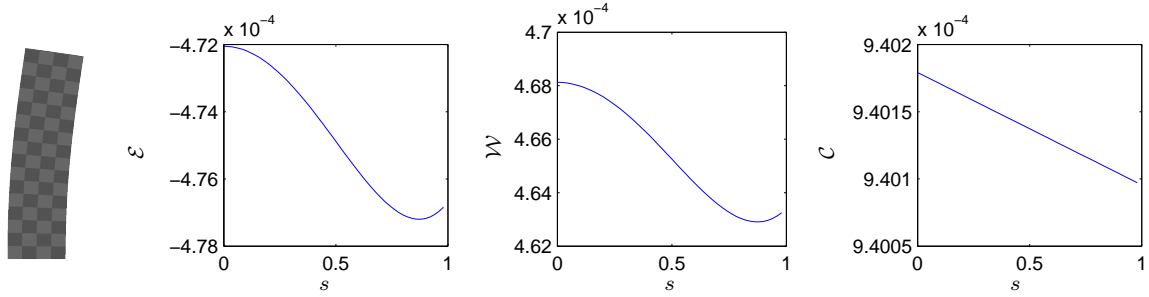


Figure 7.5: Left: Eigendisplacement of a symmetrically compressed beam (thickness $t = 0.25$) corresponding to the negative eigenvalue of the free energy Hessian. Right: Energetic changes for the perturbation of the symmetric compression in direction of the eigendisplacement. The coordinate s indicates the perturbation strength, and $s = 0$ corresponds to the symmetric deformation.

the result of linearised elasticity, apparently stops existing at \hat{t} . Indeed, for symmetry reasons we know that the symmetric deformation between buckling to the left and to the right must be a critical point of $\mathcal{E}[\mathcal{O}, \cdot]$. This deformation is readily obtained by a simple Newton iteration to find the zero of the derivative of $\mathcal{E}[\mathcal{O}, \cdot]$. The corresponding stiffness operator, that is, the Hessian of $\mathcal{E}[\mathcal{O}, \cdot]$ at this symmetric deformation then is indefinite and has a negative eigenvalue, classifying the symmetric deformation as a saddle point of the free energy. For the bar of thickness $t = 0.25$, the eigendisplacement belonging to the negative eigenvalue is shown in Figure 7.5, as well as the free energy decrease along this direction. The eigendisplacement can easily be recognised as a (linearised) bending deformation.

It is not necessarily true that the state which corresponds to the equilibrium deformation in linearised elasticity does not persist in parallel to other (local) equilibrium states. Figure 7.6 shows an example where this is not the case. As it occurred during one of our shape optimisations, it additionally serves to illustrate two particular problems we are facing during computations with nonlinear elasticity.

In Figure 7.6, we consider the structure shown top left. It is fixed at the bottom and subjected to a surface load at its top. In fact, the structure represents an intermediate result of a phase field optimisation and has to be interpreted as in Figure 7.2: White areas correspond to stiff material while black regions represent a material which is weaker by several orders of magnitude. Grey levels indicate some intermediate stiffness.

If we gradually increase the magnitude of the surface load, then for some magnitudes we can numerically detect two simultaneously existing local equilibria, that is, local minimisers ϕ of the total free energy \mathcal{E} . One of them corresponds to the equilibrium of linearised elasticity, the other one can only occur in nonlinear elasticity. This existence of multiple local equilibria poses a serious problem to shape optimisation algorithms: In general, we will only be able to detect a local minimum of the free energy \mathcal{E} and thus a locally stable deformation, not realising that the global minimum is actually different. In order to obtain the two stable deformations from Figure 7.6, we employed a homo-

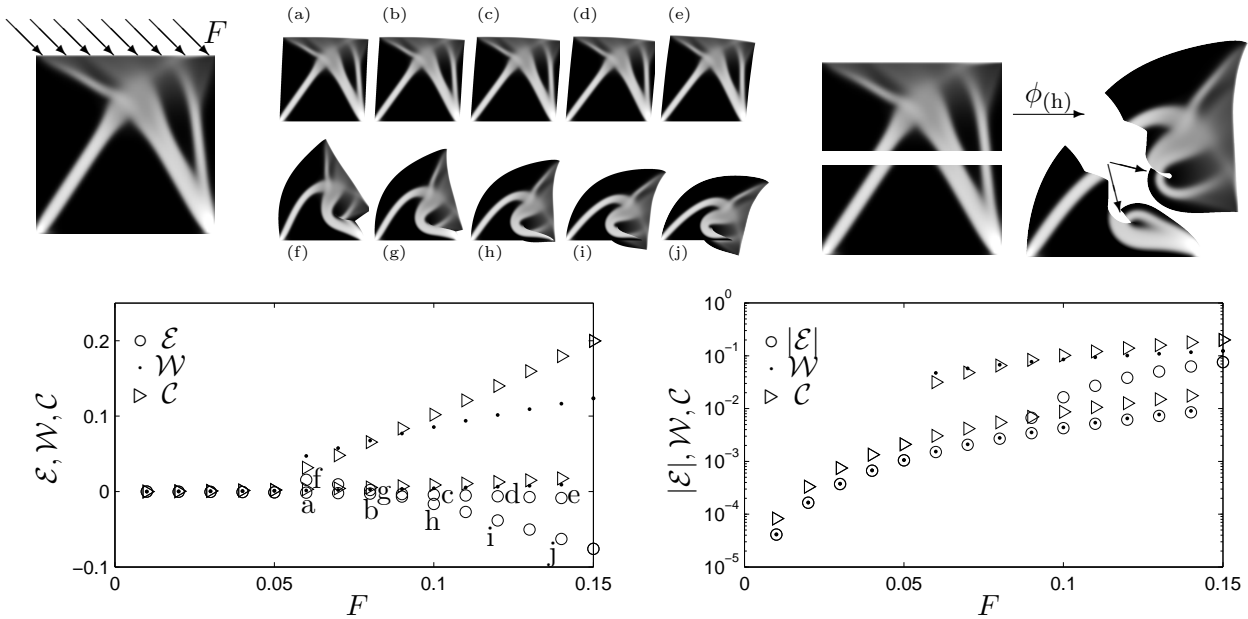


Figure 7.6: Top: Sketch of the load carried by the object (left) and resulting equilibrium deformations (from (a) to (e) and from (f) to (j)). The magnitude of the load is increased from (a) to (e) and from (f) to (j); top and bottom row show simultaneously existing local equilibria. The rightmost figure shows deformation (h) applied to the bottom and the top half of the object; the region of material interpenetration is marked by arrows. Bottom: $\mathcal{E}[\mathcal{O}, \phi]$, $\mathcal{W}[\mathcal{O}, \phi]$, and $\mathcal{C}[\phi]$ for the equilibrium deformations as a function of the load magnitude (the second graph shows a logarithmic plot).

topy method, that is, starting from the deformation of linearised elasticity, we gradually increase the load and compute the corresponding deformations using the previous deformation as initial guess. At the highest load, the minimisation of \mathcal{E} suddenly (probably due to some small perturbation) detects the different equilibrium state which we then take as initial guess to compute the second equilibrium deformations for smaller surface loads (for the actual numerical algorithm see Section 7.4).

The second difficulty associated with the use of nonlinear elasticity and illustrated by the above example is the following. A closer look reveals that the strong deformations in Figure 7.6 are, strictly speaking, unphysical: There is material interpenetration at the bottom left; one part of the structure is shifted on top of another (Figure 7.6, top right). In fact, the obtained deformation is a particular instance of the example from [12] of a (even locally) non-invertible deformation of the unit disc in \mathbb{R}^2 , given in polar coordinates (r, θ) by

$$(r, \theta) \mapsto (r, 2\theta).$$

In order to prevent such non-invertible deformations, the deformed objects can be embedded in a large, surrounding matrix of very weak material for which then Dirichlet

boundary conditions are prescribed that are consistent with globally invertible deformations (see Section 3.1 and [12]). We did not do so here since the computational effort rapidly becomes enormous.

Actually, local equilibria of the above type can also be found for the compressed vertical bar so that, formally, there are also parallelly existing local minimisers of the free energy \mathcal{E} (apart from the trivial additional state of buckling to the other side). An interesting example with physically meaningful equilibria is given in Section 7.3.1 in the context of analysing the existence of optimal shapes.

7.3 Model analysis

This section is devoted to the study of the existence problem associated with our shape or phase field optimisation. We will first prove existence of optimising phase fields, where in case of multiple global equilibrium deformations we always choose the one with least compliance. The situation when choosing a different equilibrium deformation is examined afterwards, and we will finally analyse the model behaviour as the phase field parameter ε approaches zero.

7.3.1 Existence of minimisers

We aim to establish the existence of a phase field u that minimises $\mathcal{J}_W^\varepsilon[u, \phi]$ or alternatively $\mathcal{J}_C^\varepsilon[u, \phi]$ under the constraint that ϕ minimises $\mathcal{E}[u, \phi]$. For the sake of clarity, we will proceed in small steps and first prove some properties of the functional \mathcal{E} . In the following, we will always assume $d \in \{2, 3\}$ and $\Omega \subset \mathbb{R}^d$ to be bounded, open, and connected with Lipschitz boundary without explicitly mentioning it.

Theorem 10 (Existence of equilibrium deformations). *Let $u \in W^{1,2}(\Omega)$, and let $\chi_{\mathcal{O}}[u]$ be non-negative. If $W : \mathbb{R}^{d \times d} \rightarrow \mathbb{R}$ is polyconvex with $W(A) \geq C_1 \|A\|_F^p - C_2$, $p \geq d$, and $F \in L^{p'}(\Gamma_N)$ for $1 = \frac{1}{p} + \frac{1}{p'}$, then the variational problem $\min_{\phi} \mathcal{E}[u, \phi]$ admits a minimiser in $\{\phi \in W^{1,p}(\Omega) : \phi|_{\Gamma_D} = \text{id}\}$.*

Proof. This is a direct consequence of Theorem 1 since $((1-\delta)\chi_{\mathcal{O}}[u] + \delta)W(\cdot)$ is obviously also polyconvex and satisfies a p -growth condition. \square

In the previous theorem, the existence of equilibrium deformations heavily relies on the weak lower semi-continuity of the energy $\mathcal{W}[u, \cdot]$ for a fixed phase field u . However, during shape optimisation we vary the phase field so that we will later need weak lower semi-continuity with respect to both, the deformation and the phase field, as provided by the following lemma.

Lemma 11. *Let $\chi_{\mathcal{O}}[u]$ be non-negative and continuous in u and $W : \mathbb{R}^{d \times d} \rightarrow \mathbb{R}$ bounded from below and polyconvex, then $\mathcal{W}[u, \phi]$ and thus $\mathcal{E}[u, \phi]$ are sequentially lower semi-continuous along sequences $(u_i, \phi_i)_{i \in \mathbb{N}}$ with $u_i \rightarrow u$ in $L^1(\Omega)$ and $(\mathcal{D}\phi_i, \text{cof}\mathcal{D}\phi_i, \det\mathcal{D}\phi_i) \rightarrow (\mathcal{D}\phi, \text{cof}\mathcal{D}\phi, \det\mathcal{D}\phi)$ in $L^p(\Omega) \times L^q(\Omega) \times L^r(\Omega)$ for $p, q, r \geq 1$.*

Proof. Without loss of generality we may assume $W > 0$. Let $\underline{\mathcal{W}} := \liminf_{i \rightarrow \infty} \mathcal{W}[u_i, \phi_i]$, then upon extracting a subsequence and after reindexing, we may assume

$$\lim_{i \rightarrow \infty} \mathcal{W}[u_i, \phi_i] = \underline{\mathcal{W}}.$$

By Mazur's lemma there are $N_k \in \mathbb{N}$ and $\lambda_i^k \in [0, 1]$ with $\sum_{i=k}^{N_k} \lambda_i^k = 1$ and

$$\sum_{i=k}^{N_k} \lambda_i^k (\mathcal{D}\phi_i, \text{cof}\mathcal{D}\phi_i, \det\mathcal{D}\phi_i) \rightarrow (\mathcal{D}\phi, \text{cof}\mathcal{D}\phi, \det\mathcal{D}\phi)$$

strongly in $L^p(\Omega) \times L^q(\Omega) \times L^r(\Omega)$. Also, due to the polyconvexity of W , we may write $W(\mathcal{D}\phi_i) = \bar{W}(\mathcal{D}\phi_i, \text{cof}\mathcal{D}\phi_i, \det\mathcal{D}\phi_i)$ for a convex function \bar{W} . Then

$$\begin{aligned} \underline{\mathcal{W}} &= \liminf_{k \rightarrow \infty} \sum_{i=k}^{N_k} \lambda_i^k \int_{\Omega} \chi_{\mathcal{O}}[u_i] \bar{W}(\mathcal{D}\phi_i, \text{cof}\mathcal{D}\phi_i, \det\mathcal{D}\phi_i) \, dx \\ &\geq \liminf_{k \rightarrow \infty} \sum_{i=k}^{N_k} \lambda_i^k \int_{\Omega} \left(\inf_{j=k, \dots, N_k} \chi_{\mathcal{O}}[u_j] \right) \bar{W}(\mathcal{D}\phi_i, \text{cof}\mathcal{D}\phi_i, \det\mathcal{D}\phi_i) \, dx \\ &= \liminf_{k \rightarrow \infty} \int_{\Omega} \left(\inf_{j=k, \dots, N_k} \chi_{\mathcal{O}}[u_j] \right) \sum_{i=k}^{N_k} \lambda_i^k \bar{W}(\mathcal{D}\phi_i, \text{cof}\mathcal{D}\phi_i, \det\mathcal{D}\phi_i) \, dx \\ &\geq \liminf_{k \rightarrow \infty} \int_{\Omega} \left(\inf_{j=k, \dots, N_k} \chi_{\mathcal{O}}[u_j] \right) \bar{W} \left(\sum_{i=k}^{N_k} \lambda_i^k (\mathcal{D}\phi_i, \text{cof}\mathcal{D}\phi_i, \det\mathcal{D}\phi_i) \right) \, dx \\ &\geq \int_{\Omega} \liminf_{k \rightarrow \infty} \left(\inf_{j=k, \dots, N_k} \chi_{\mathcal{O}}[u_j] \right) \bar{W} \left(\sum_{i=k}^{N_k} \lambda_i^k (\mathcal{D}\phi_i, \text{cof}\mathcal{D}\phi_i, \det\mathcal{D}\phi_i) \right) \, dx, \end{aligned}$$

where we have exploited the convexity of \bar{W} and applied Fatou's lemma. We finally deduce

$$\begin{aligned} \underline{\mathcal{W}} &\geq \int_{\Omega} \left(\liminf_{k \rightarrow \infty} \inf_{j=k, \dots, N_k} \chi_{\mathcal{O}}[u_j] \right) \left(\liminf_{k \rightarrow \infty} \bar{W} \left(\sum_{i=k}^{N_k} \lambda_i^k (\mathcal{D}\phi_i, \text{cof}\mathcal{D}\phi_i, \det\mathcal{D}\phi_i) \right) \right) \, dx \\ &= \int_{\Omega} \chi_{\mathcal{O}}[u] \bar{W} \left(\liminf_{k \rightarrow \infty} \sum_{i=k}^{N_k} \lambda_i^k (\mathcal{D}\phi_i, \text{cof}\mathcal{D}\phi_i, \det\mathcal{D}\phi_i) \right) \, dx = \mathcal{W}[u, \phi], \end{aligned}$$

where we have used the pointwise convergence of $\sum_{i=k}^{N_k} \lambda_i^k (\mathcal{D}\phi_i, \text{cof}\mathcal{D}\phi_i, \det\mathcal{D}\phi_i)$ and u_k (and thus $\chi_{\mathcal{O}}[u_k]$) almost everywhere (possibly after extracting a subsequence) and the lower semi-continuity of \bar{W} . \square

Lemma 12. *Let $0 \leq \chi_{\mathcal{O}}[u] \leq 1$ be continuous in u , and let $W : \mathbb{R}^{d \times d} \rightarrow \mathbb{R}$ be polyconvex with $W(A) \geq C_1 \|A\|_F^p - C_2$, $p > d$, and $F \in L^{p'}(\Gamma_N)$ for $1 = \frac{1}{p} + \frac{1}{p'}$. If $u_i \rightarrow u$ in $L^1(\Omega)$, then*

$$\Gamma - \lim_{i \rightarrow \infty} \mathcal{E}[u_i, \cdot] = \mathcal{E}[u, \cdot]$$

with respect to the weak $W^{1,p}(\Omega)$ -topology.

Proof. Since the boundary integral $-\mathcal{C}[\cdot]$ is just a continuous perturbation, we need to show Γ -convergence of $\mathcal{W}[u_i, \cdot]$ only.

Let $\phi_i \rightharpoonup \phi$ in $W^{1,p}(\Omega)$ with $\limsup_{i \rightarrow \infty} \mathcal{W}[u_i, \phi_i] < \infty$. By the growth conditions on W we deduce the boundedness of $(\text{cof} \mathcal{D}\phi_i, \det \mathcal{D}\phi_i)$ in $L^{p/(d-1)}(\Omega) \times L^{p/d}(\Omega)$ and thus—due to the reflexivity of the Lebesgue spaces—the weak convergence of a subsequence. Then, as in Section 3.1, we can apply Ball’s compensated compactness result to obtain $(\mathcal{D}\phi_i, \text{cof} \mathcal{D}\phi_i, \det \mathcal{D}\phi_i) \rightharpoonup (\mathcal{D}\phi, \text{cof} \mathcal{D}\phi, \det \mathcal{D}\phi)$ in $L^p(\Omega) \times L^{p/(d-1)}(\Omega) \times L^{p/d}(\Omega)$ [11]. The previous lemma then yields the lim inf-inequality.

For the lim sup-inequality, note that $\mathcal{W}[u_i, \phi] \rightarrow \mathcal{W}[u, \phi]$. Otherwise there would be a $\rho > 0$ and a subsequence u_j , $j \in J \subset \mathbb{N}$, such that $|\mathcal{W}[u_j, \phi] - \mathcal{W}[u, \phi]| > \rho$ for all $j \in J$. Since $u_j \rightarrow u$ in $L^1(\Omega)$, we can then extract a further subsequence u_j , $j \in \hat{J} \subset J$, such that $u_j \rightarrow u$ pointwise almost everywhere as $j \rightarrow \infty$ in \hat{J} . The integrand of $\mathcal{W}[u_j, \phi]$ is bounded above by $(1 + \delta)W(\mathcal{D}\phi)$ and converges pointwise against $\chi_{\mathcal{O}}[u]W(\mathcal{D}\phi)$. By the dominated convergence theorem, we obtain $\mathcal{W}[u_j, \phi] \rightarrow \mathcal{W}[u, \phi]$ as $j \rightarrow \infty$ in \hat{J} , which is a contradiction. Hence, for the recovery sequence $\phi_i = \phi$, $i \in \mathbb{N}$, we obtain $\limsup_{i \rightarrow \infty} \mathcal{W}[u_i, \phi_i] = \mathcal{W}[u, \phi]$ which proves the lim sup-inequality. \square

Let us denote by $\mathbf{m}[u]$ the set of minimisers from Theorem 10. We now prove the existence of minimising phase fields u for our constrained minimisation problem.

Theorem 13 (Existence of optimal phase fields). *Let $0 \leq \chi_{\mathcal{O}}[u] \leq 1$ be continuous in u , $W : \mathbb{R}^{d \times d} \rightarrow \mathbb{R}$ be polyconvex with $W(A) \geq C_1 \|A\|_F^p - C_2$, $p > d$, and let $F \in L^{p'}(\Gamma_N)$ for $1 = \frac{1}{p} + \frac{1}{p'}$. Then the variational problems $\min_u \mathcal{J}_{\mathcal{W}}^\varepsilon[u, \phi]$ and $\min_u \mathcal{J}_{\mathcal{C}}^\varepsilon[u, \phi]$ admit minimisers under the constraint $\phi \in \mathbf{m}[u]$.*

Proof. At first, note that for any $u \in W^{1,2}(\Omega)$ we have $\mathcal{E}[u, \text{id}] = 0$ and thus $\mathcal{E}[u, \phi] \leq 0$ for all $\phi \in \mathbf{m}[u]$. We deduce $\mathcal{W}[u, \phi] \leq \mathcal{C}[\phi]$ for all $u \in W^{1,2}(\Omega)$ and $\phi \in \mathbf{m}[u]$.

Apparently, $\mathcal{J}_{\mathcal{W}}^\varepsilon$ and $\mathcal{J}_{\mathcal{C}}^\varepsilon$ are bounded from below by $-C_2|\Omega|$, and they are not constantly $+\infty$ since for $u = 0$ and $\phi \in \mathbf{m}[u]$, $\mathcal{W}[u, \phi]$ and $\mathcal{C}[\phi]$ are bounded (see proofs of Theorems 1 and 10). We consider a minimising sequence u_i , $i \in \mathbb{N}$ (a different one for $\mathcal{J}_{\mathcal{W}}^\varepsilon$ and $\mathcal{J}_{\mathcal{C}}^\varepsilon$, respectively). Due to the weak $W^{1,2}(\Omega)$ -coercivity of $\mathcal{J}_{\mathcal{W}}^\varepsilon$ and $\mathcal{J}_{\mathcal{C}}^\varepsilon$ with respect to the phase field (by virtue of the regularisation $\mathcal{L}_{\text{MM}}^\varepsilon[u]$), there is $u \in W^{1,2}(\Omega)$ with $u_i \rightharpoonup u$ as $i \rightarrow \infty$ (after extraction of a subsequence).

Let $\phi[u_i] \in \mathbf{m}[u_i]$ denote the sequence of deformations belonging to the minimising sequence u_i . Due to the growth conditions on W as well as $\mathcal{J}_{\mathcal{C}}^\varepsilon \geq \mathcal{J}_{\mathcal{W}}^\varepsilon \geq \mathcal{W}$ and the Poincaré inequality, we know that $\phi[u_i]$ must be uniformly bounded in $W^{1,p}(\Omega)$. Hence,

by the reflexivity of $W^{1,p}(\Omega)$, there is $\phi \in W^{1,p}(\Omega)$ with $\phi[u_i] \rightharpoonup \phi$ (after extracting a subsequence). Since $\mathcal{E}[u_i, \cdot]$ is equi-mildly coercive (see proofs of Theorems 1 and 10), Lemma 12 and Theorem 3 imply $\phi \in \mathbf{m}[u]$. Here, note that the Γ -limit is consistent with the Dirichlet boundary conditions at Γ_D .

Finally, $\mathcal{J}_W^\varepsilon[u_i, \phi[u_i]]$ and $\mathcal{J}_C^\varepsilon[u_i, \phi[u_i]]$ are sequentially weakly lower semi-continuous as $u_i \rightharpoonup u$ in $W^{1,2}(\Omega)$ and $\phi[u_i] \rightharpoonup \phi$ in $W^{1,p}(\Omega)$: The lower semi-continuity of $\mathcal{L}_{\text{MM}}^\varepsilon[u_i]$ and $\mathcal{V}[u_i]$ is obvious as their integrands are convex in ∇u and continuous in u . By a trace theorem, $\phi[u_i] \rightarrow \phi$ strongly in $L^p(\Gamma_N)$ so that $\mathcal{C}[\phi[u_i]]$ is also lower semi-continuous. Furthermore, as in the proof of the previous lemma, we may assume $(\mathcal{D}\phi[u_i], \text{cof}\mathcal{D}\phi[u_i], \det\mathcal{D}\phi[u_i]) \rightharpoonup (\mathcal{D}\phi, \text{cof}\mathcal{D}\phi, \det\mathcal{D}\phi)$ so that the lower semi-continuity of $\mathcal{W}[u_i, \phi[u_i]]$ follows from Lemma 11.

From the above, $\mathcal{J}_W^\varepsilon[u, \phi] \leq \liminf_{i \rightarrow \infty} \mathcal{J}_W^\varepsilon[u_i, \phi[u_i]]$ ($\mathcal{J}_C^\varepsilon[u, \phi] \leq \liminf_{i \rightarrow \infty} \mathcal{J}_C^\varepsilon[u_i, \phi[u_i]]$), respectively), and u is a minimiser. \square

Remark. Since the set $\{u \in W^{1,2}(\Omega) : -1 \leq u \leq 1\}$ is closed in $W^{1,2}(\Omega)$ with respect to the weak topology, the results still hold if we restrict the phase field u to take values in $[-1, 1]$ so that for $\chi_{\mathcal{O}}[u]$ we only have to require $\chi_{\mathcal{O}}[\cdot] : [-1, 1] \rightarrow [0, 1]$.

7.3.2 Non-existence of minimisers for worst case deformations

The above result only states that there is one equilibrium deformation $\phi[u] \in \mathbf{m}[u]$ such that $\mathcal{J}^\varepsilon[u, \phi[u]]$ is minimal. There might be more equilibrium deformations $\phi \in \mathbf{m}[u]$ for which $\mathcal{J}^\varepsilon[u, \phi] > \mathcal{J}^\varepsilon[u, \phi[u]]$. Such deformations may be seen as worst case scenarios: They represent possible equilibrium configurations with stronger strains. For this reason, it might be more interesting to actually consider (for $\mathcal{G} = \mathcal{W}$ or $\mathcal{G} = \mathcal{C}$) the objective functional $\overline{\mathcal{J}}_{\mathcal{G}}^\varepsilon[u] := \sup_{\phi \in \mathbf{m}[u]} \mathcal{J}_{\mathcal{G}}^\varepsilon[u, \phi]$. However, minimisers for $\overline{\mathcal{J}}_{\mathcal{G}}^\varepsilon$ seem to generally not exist as the following example illustrates.

We would like to optimise the structure in Figure 7.7, left. It is clamped at its bottom and subjected to a surface load from the top. Its right-most edge can move freely in vertical direction, but is fixed in horizontal direction. Furthermore, its vertical pillar is restricted to stay left of the line indicated in the sketch (which may be interpreted as a wall that cannot be penetrated). Instead of optimising the structure within the set of all possible shapes, we will only consider a simple, one-dimensional subset which is generated by eroding the original shape depicted in the sketch. That is, we try to find just the optimal thickness of the object.

The top and bottom row of the table in Figure 7.7 depict two different equilibrium deformations for the prescribed loading (that is, stable deformations which are local minimisers of \mathcal{E}), where each column belongs to a different object thickness: From left to right, the original shape (whose vertical pillar has a width of six length units) is eroded by zero up to four length units (in the computations, one length unit actually corresponds to a pixel). The corresponding total free energy \mathcal{E} as well as the stored elastic energy \mathcal{W} and the change in external potential \mathcal{C} are shown in the right graph,

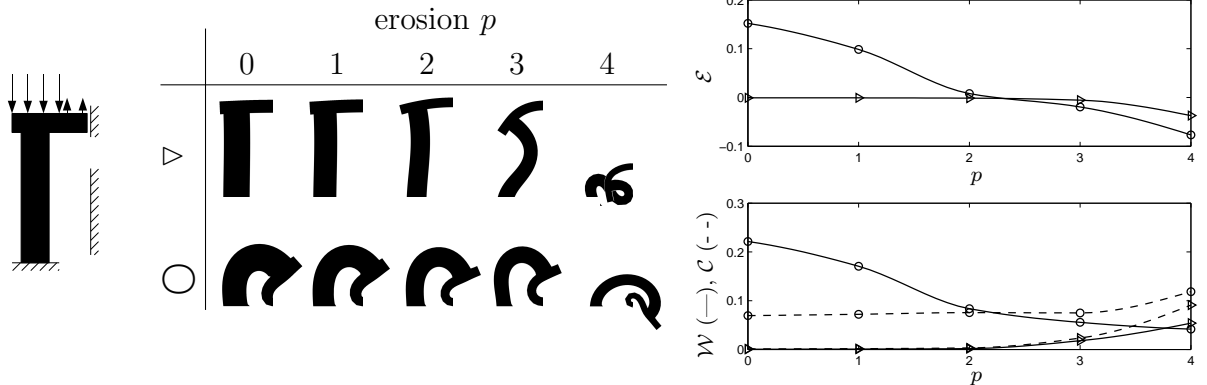


Figure 7.7: Left: Sketch of the shape optimisation problem under consideration. Middle: Equilibrium deformations for shapes of different thickness, starting from a reference shape on the left and gradually eroding it to the right. The top and bottom row represent two different equilibria. Right: Energy components of the different configurations.

where triangles and circles represent the energies of the configurations in the upper and lower table row, respectively. For ease of reference, we shall denote the deformations by $\phi_{\triangleright p}$ and $\phi_{\circ p}$, $0 \leq p \leq 4$.

Assume, we start the optimisation from the thickest shape (left end of table and graph). The equilibrium deformation which globally minimises \mathcal{E} is apparently given by $\phi_{\triangleright 0}$ (compare Figure 7.7, right, and the qualitative sketch in Figure 7.8). Assume that the volume penalty parameter ν is chosen large enough so that the objective functional $\overline{\mathcal{J}}_G^\mathcal{E}$ decreases for increasing erosion p . Then $\phi_{\triangleright p}$ stays the global equilibrium deformation until a point $p = \hat{p}$ between 2 and 3, where suddenly $\phi_{\circ p}$ takes over as the global equilibrium deformation. At this point, the objective functional jumps discontinuously to a higher value, since \mathcal{W} as well as \mathcal{C} are larger for the deformations $\phi_{\circ p}$ than $\phi_{\triangleright p}$. Hence, if the weight ν of the volume term is chosen such that—neglecting the existence of the equilibrium deformations $\phi_{\circ p}$ —the optimal thickness would lie exactly at \hat{p} or slightly beyond it, then there will be no minimiser: From the left we can get arbitrarily close to \hat{p} and thus the objective functional gets arbitrarily close to its infimum value, but we cannot reach it since at \hat{p} , the cost functional suddenly jumps up.

For the sake of completeness, let us also briefly explain the idea behind choosing the above example. We seek for a configuration with two simultaneously existing equilibrium deformations such as a compressed bar or pillar which can buckle to its left or right side. One configuration should be initially preferred (that is, be the global minimiser of \mathcal{E} in the case of rather thick and thus stiff structures), while the other should take over at some point if the structure becomes less stiff and is deformed more strongly. The initial preference for rightward buckling is achieved by adding the slight upward traction at the top right of the shape. If the pillar buckles so strongly that it touches the wall on

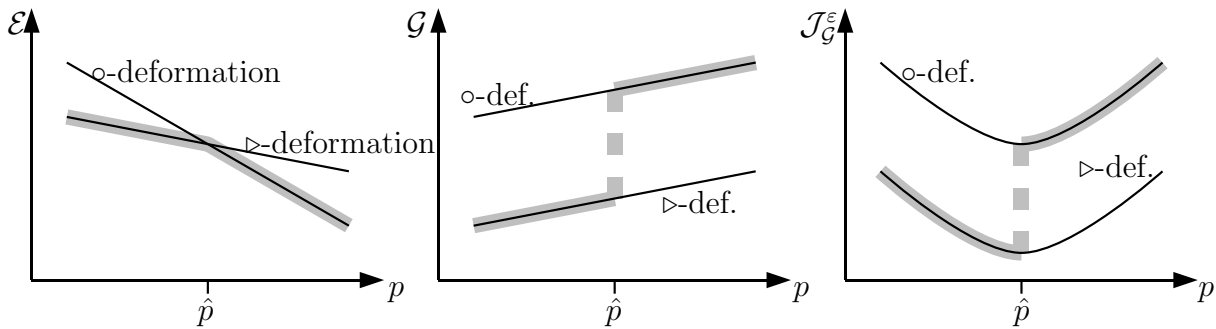


Figure 7.8: Qualitative sketch of the energies belonging to the \circ - and \triangleright -deformations from Figure 7.7. \mathcal{G} stands for \mathcal{W} or \mathcal{C} . The curves belonging to the global equilibrium deformation (the global minimiser of \mathcal{E}) are highlighted in grey. At $p = \hat{p}$, the global equilibrium deformation switches from $\phi_{\triangleright p}$ to $\phi_{\circ p}$, which is associated with a jump in \mathcal{G} and a corresponding jump in $\mathcal{J}_{\mathcal{G}}^{\epsilon}$ (thick grey curve in the rightmost graph).



Figure 7.9: Local equilibrium deformation for the mechanical problem in Figure 7.7 which reverses the orientation of the crossbeam and cannot be reached starting from the configuration in Figure 7.7, left.

the right, the structure stiffens (due to the additional support by the wall), and hence the leftward buckling will at some point yield less free energy and become the global minimiser of \mathcal{E} .

A nice peculiarity of the above example is that the equilibrium deformation $\phi_{\circ p}$ initially yields a positive value of \mathcal{E} , or in different words, that the stored elastic energy \mathcal{W} is larger than the external potential change \mathcal{C} . This is associated with a self-locking mechanism: Even without loads, a deformation of the type $\phi_{\circ p}$ would be stable, and the object cannot be deformed into a less strained state without intermediately increasing the strain.

Having presented this example, the question arises naturally whether one should actually go even further and consider \mathcal{W} and \mathcal{C} for all local equilibrium deformations, that is, extend the set $\mathfrak{m}[u]$ to the set of all local minimisers of $\mathcal{E}[u, \cdot]$. However, this complicates the system even further, and one would also have to pay attention to exclude unphysical states that, for example, imply a local reversion of orientation such as shown for the above example in Figure 7.9.

Remark. In Section 7.2.1, we proposed two possible definitions for the compliance in the nonlinear elasticity setting: The internal elastic energy \mathcal{W} and the change of external potential \mathcal{C} . Both are consistent with the compliance definition for linearised

elasticity, and their minimisation yields optimally rigid (in the case of \mathcal{C}) or least strained structures (for minimising \mathcal{W}). However, there is a third possibility which reduces to the standard notion of compliance in linearised elasticity. Indeed, we might also choose to minimise the dissipation associated with the transition from the unstressed state to the equilibrium deformation, $-\mathcal{E} = \mathcal{C} - \mathcal{W}$. Since by definition, for a given phase field u we have $\sup_{\phi \in \mathfrak{m}[u]}(-\mathcal{E}[u, \phi]) = \inf_{\phi \in \mathfrak{m}[u]}(-\mathcal{E}[u, \phi]) = -\inf_{\phi} \mathcal{E}[u, \phi]$, we can establish existence of minimisers for

$$\mathcal{J}_{-\mathcal{E}}^\varepsilon[u, \phi] = -2\mathcal{E}[u, \phi] + \nu\mathcal{V}[u] + \eta\mathcal{L}_{\text{MM}}^\varepsilon[u]$$

under the constraint $\phi \in \mathfrak{m}[u]$ even in the worst case.

7.3.3 Model behaviour for phase field parameter $\varepsilon \rightarrow 0$

So far, we have assumed the phase field parameter ε to be fixed. However, we are actually interested in the limit case of sharp interfaces, which we hope to retrieve as we let $\varepsilon \rightarrow 0$. Unfortunately, the non-uniqueness of the equilibrium deformation prevents us from proving a general Γ -convergence result: It might theoretically happen that—as ε reaches zero and the phase field interface gets ultimately sharp—suddenly an additional equilibrium deformation occurs which results in a sudden increase or decrease of the objective functional value. For this reason, we can only state the following two weaker results.

Let us define $\underline{\mathcal{J}}_{\mathcal{G}}^\varepsilon[u] := \inf_{\phi \in \mathfrak{m}[u]} \mathcal{J}_{\mathcal{G}}^\varepsilon[u, \phi]$ and $\overline{\mathcal{J}}_{\mathcal{G}}^\varepsilon[u] := \sup_{\phi \in \mathfrak{m}[u]} \mathcal{J}_{\mathcal{G}}^\varepsilon[u, \phi]$ for $\mathcal{G} = \mathcal{W}$ or $\mathcal{G} = \mathcal{C}$. Furthermore, define

$$\begin{aligned} \underline{\mathcal{J}}_{\mathcal{G}}^0[u] &:= \begin{cases} \inf_{\phi \in \mathfrak{m}[u]} \mathcal{G}[u, \phi] + \nu\mathcal{V}[u] + \frac{\eta}{2}|u|_{\text{TV}(\Omega)}, & u : \Omega \rightarrow \{-1, 1\} \\ \infty, & \text{else} \end{cases}, \\ \overline{\mathcal{J}}_{\mathcal{G}}^0[u] &:= \begin{cases} \sup_{\phi \in \mathfrak{m}[u]} \mathcal{G}[u, \phi] + \nu\mathcal{V}[u] + \frac{\eta}{2}|u|_{\text{TV}(\Omega)}, & u : \Omega \rightarrow \{-1, 1\} \\ \infty, & \text{else} \end{cases}, \end{aligned}$$

where $|\cdot|_{\text{TV}(\Omega)}$ denotes the total variation.

Theorem 14. *Under the conditions of Theorem 13, we have*

$$\Gamma - \liminf_{\varepsilon \rightarrow 0} \underline{\mathcal{J}}_{\mathcal{G}}^\varepsilon \geq \underline{\mathcal{J}}_{\mathcal{G}}^0$$

with respect to the $L^1(\Omega)$ -topology.

Proof. Let $u_\varepsilon \rightarrow u$ in $L^1(\Omega)$ as $\varepsilon \rightarrow 0$, then obviously $\mathcal{V}[u_\varepsilon] \rightarrow \mathcal{V}[u]$. Furthermore,

$$\liminf_{\varepsilon \rightarrow 0} \mathcal{L}_{\text{MM}}^\varepsilon[u_\varepsilon] \geq \begin{cases} \frac{1}{2}|u|_{\text{TV}(\Omega)}, & u : \Omega \rightarrow \{-1, 1\} \\ \infty, & \text{else} \end{cases}$$

by Theorem 4. Finally, either $\liminf_{\varepsilon \rightarrow 0} \inf_{\phi \in \mathbf{m}[u_\varepsilon]} \mathcal{G}[u_\varepsilon, \phi] = \infty$, in which case there is nothing left to prove, or there is a sequence $(\varepsilon_i)_{i \in \mathbb{N}}$ with $\varepsilon_i \rightarrow 0$ as $i \rightarrow \infty$ and a sequence ϕ_i with $\phi_i \in \mathbf{m}[u_{\varepsilon_i}]$ such that

$$\lim_{i \rightarrow \infty} \mathcal{G}[u_{\varepsilon_i}, \phi_i] = \liminf_{\varepsilon \rightarrow 0} \inf_{\phi \in \mathbf{m}[u_\varepsilon]} \mathcal{G}[u_\varepsilon, \phi] < \infty.$$

As in the proof of Theorem 13, we use the growth conditions on W and Theorems 12 and 3 to deduce that—for a subsequence— $\phi_i \rightarrow \phi[u]$ in $W^{1,p}(\Omega)$ and $\mathcal{E}[u_\varepsilon, \phi_\varepsilon] \rightarrow \mathcal{E}[u, \phi[u]]$ for some $\phi[u] \in \mathbf{m}[u]$. Also, $\mathcal{C}[\phi_\varepsilon] \rightarrow \mathcal{C}[\phi[u]]$ due to the continuity of \mathcal{C} and thus also $\mathcal{W}[u_\varepsilon, \phi_\varepsilon] = \mathcal{E}[u_\varepsilon, \phi_\varepsilon] + \mathcal{C}[\phi_\varepsilon] \rightarrow \mathcal{E}[u, \phi[u]] + \mathcal{C}[\phi[u]] = \mathcal{W}[u, \phi[u]]$ so that

$$\liminf_{\varepsilon \rightarrow 0} \inf_{\phi \in \mathbf{m}[u_\varepsilon]} \mathcal{G}[u_\varepsilon, \phi] = \lim_{i \rightarrow \infty} \mathcal{G}[u_{\varepsilon_i}, \phi_i] = \mathcal{G}[u, \phi[u]] \geq \inf_{\phi \in \mathbf{m}[u]} \mathcal{G}[u, \phi],$$

which altogether yields the desired result. \square

Theorem 15. *Under the conditions of Theorem 13, we have*

$$\Gamma - \limsup_{\varepsilon \rightarrow 0} \overline{\mathcal{J}_G^\varepsilon} \leq \overline{\mathcal{J}_G^0}$$

with respect to the $L^1(\Omega)$ -topology.

Proof. Let $u_\varepsilon \rightarrow u$ in $L^1(\Omega)$ be a recovery sequence with respect to the Γ -convergence of $\mathcal{L}_{\text{MM}}^\varepsilon$. As before, we have $\mathcal{V}[u_\varepsilon] \rightarrow \mathcal{V}[u]$. Furthermore,

$$\limsup_{\varepsilon \rightarrow 0} \mathcal{L}_{\text{MM}}^\varepsilon[u_\varepsilon] \leq \begin{cases} \frac{1}{2}|u|_{\text{TV}(\Omega)}, & u : \Omega \rightarrow \{-1, 1\} \\ \infty, & \text{else} \end{cases}$$

by Theorem 4. Finally, as in the previous proof, there are sequences ε_i and ϕ_i with $\varepsilon_i \rightarrow 0$, $\phi_i \in \mathbf{m}[u_{\varepsilon_i}]$, and $\phi_i \rightarrow \phi[u]$ for some $\phi[u] \in \mathbf{m}[u]$ such that

$$\limsup_{\varepsilon \rightarrow 0} \sup_{\phi \in \mathbf{m}[u_\varepsilon]} \mathcal{G}[u_\varepsilon, \phi] = \lim_{i \rightarrow \infty} \mathcal{G}[u_{\varepsilon_i}, \phi_i] = \mathcal{G}[u, \phi[u]] \leq \sup_{\phi \in \mathbf{m}[u]} \mathcal{G}[u, \phi],$$

which concludes the proof. \square

If for a given phase field u there is just one single unique equilibrium deformation, then, obviously, $\overline{\mathcal{J}_G^0}[u] = \underline{\mathcal{J}_G^0}[u]$, which implies following corollary.

Corollary 16. *Let the conditions of Theorem 13 hold, and let $u : \Omega \rightarrow \mathbb{R}$ be given. If the equilibrium deformation is unique, that is, $\mathbf{m}[u] = \{\phi[u]\}$ for a $\phi[u] \in W^{1,p}(\Omega)$, then the Γ -limit of $\overline{\mathcal{J}_G^\varepsilon}$ and $\underline{\mathcal{J}_G^\varepsilon}$ for $\varepsilon \rightarrow 0$ with respect to the $L^1(\Omega)$ -topology is defined at u and is given by*

$$\overline{\mathcal{J}_G^0}[u] = \underline{\mathcal{J}_G^0}[u].$$

As mentioned earlier, since the equilibrium deformation in general is not unique, we cannot state a general Γ -convergence result. However, note that all the above proofs also hold with slight modifications in the case of linearised elasticity, that is, for

$$\mathcal{W}[u, \phi] = \mathcal{W}^{\text{lin}}[u, \phi] = \int_{\Omega} ((1 - \delta)\chi_{\mathcal{O}}[u] + \delta) \frac{1}{2} \mathbf{C} \epsilon[\phi - \text{id}] : \epsilon[\phi - \text{id}] \, dx$$

with a symmetric positive definite elasticity tensor \mathbf{C} and $\epsilon[v] = \frac{1}{2}(\mathcal{D}v + \mathcal{D}v^T)$. In this case we do obtain Γ -convergence of the objective functional.

Corollary 17. *For $\mathcal{W} = \mathcal{W}^{\text{lin}}$, $0 \leq \chi_{\mathcal{O}}[u] \leq 1$ continuous in u , and $F \in L^2(\Gamma_N)$ we have*

$$\Gamma - \lim_{\varepsilon \rightarrow 0} \overline{\mathcal{J}_{\mathcal{G}}^{\varepsilon}} = \Gamma - \lim_{\varepsilon \rightarrow 0} \underline{\mathcal{J}_{\mathcal{G}}^{\varepsilon}} = \overline{\mathcal{J}_{\mathcal{G}}^0} = \underline{\mathcal{J}_{\mathcal{G}}^0}$$

with respect to the $L^1(\Omega)$ -topology.

Proof. By Korn's first inequality, $\mathcal{W}^{\text{lin}}[u, \cdot]$ is coercive on $\{\phi \in W^{1,2}(\Omega) : \phi|_{\Gamma_D} = \text{id}\}$; furthermore, it is bounded so that the Lax–Milgram lemma implies the existence of a unique minimiser $\phi[u]$ of $\mathcal{E}^{\text{lin}}[u, \cdot]$ for which $2\mathcal{W}^{\text{lin}}[u, \phi[u]] = \mathcal{C}[\phi[u]]$. Hence, in this case we obtain $\overline{\mathcal{J}_{\mathcal{G}}^{\varepsilon}} = \underline{\mathcal{J}_{\mathcal{G}}^{\varepsilon}}$, $\overline{\mathcal{J}_{\mathcal{G}}^0} = \underline{\mathcal{J}_{\mathcal{G}}^0}$ and thus the desired result by Theorems 14 and 15. \square

The weak equi-coercivity of $\overline{\mathcal{J}_{\mathcal{G}}^{\varepsilon}} = \underline{\mathcal{J}_{\mathcal{G}}^{\varepsilon}}$ (see [18, Lemma 6.2]) then implies existence of minimisers for the sharp interface problem via Theorem 3.

Alternatively, instead of the internal elastic energy, $\mathcal{G} = \mathcal{W}$, or the change of external potential, $\mathcal{G} = \mathcal{C}$, we might also choose $\mathcal{G} = -\mathcal{E}$, the dissipation associated with the transition from the unstressed state to the equilibrium deformation, which in linearised elasticity exactly equals \mathcal{W} . All previous results still hold in that case, and furthermore $\overline{\mathcal{J}_{-\mathcal{E}}^{\varepsilon}} = \underline{\mathcal{J}_{-\mathcal{E}}^{\varepsilon}}$ as well as $\overline{\mathcal{J}_{-\mathcal{E}}^0} = \underline{\mathcal{J}_{-\mathcal{E}}^0}$ by definition of $\mathcal{E}[u, \phi]$ and $\mathfrak{m}[u]$. Hence, we again obtain the following.

Corollary 18. *Under the conditions of Theorem 13, we have*

$$\Gamma - \lim_{\varepsilon \rightarrow 0} \overline{\mathcal{J}_{-\mathcal{E}}^{\varepsilon}} = \Gamma - \lim_{\varepsilon \rightarrow 0} \underline{\mathcal{J}_{-\mathcal{E}}^{\varepsilon}} = \overline{\mathcal{J}_{-\mathcal{E}}^0} = \underline{\mathcal{J}_{-\mathcal{E}}^0}$$

with respect to the $L^1(\Omega)$ -topology.

As above, we can deduce the existence of minimisers for the sharp interface problem.

7.4 Numerical implementation

In the following sections, we shall first state the optimality conditions of the minimisation problem and its discretisation by finite elements. We then briefly describe the computation of equilibrium deformations via a trust region method and the optimisation for the phase field by a quasi-Newton method, embedded in a multiscale approach.

7.4.1 Optimality conditions and finite element discretisation

A necessary condition for ϕ to satisfy the constraint of static equilibrium is that it fulfils the Euler–Lagrange condition

$$0 = \langle \delta_\phi \mathcal{E}, \theta \rangle = \int_{\Omega} ((1 - \delta)\chi_{\mathcal{O}}[u] + \delta)W_{,A}(\mathcal{D}\phi) : \mathcal{D}\theta \, dx - \int_{\Gamma_N} F \cdot \theta \, da$$

for all test displacements $\theta : \Omega \rightarrow \mathbb{R}^d$ with $\theta|_{\Gamma_D} = 0$, where $\langle \delta_z \mathcal{G}, \zeta \rangle$ denotes the Gâteaux derivative of an energy \mathcal{G} with respect to z in some test direction ζ . This is the weak form of a pointwise partial differential equation constraint on ϕ . Hence, by the first order optimality conditions, the solution to our shape optimisation problem can be described as a saddle point of the Lagrange functional

$$L[u, \phi, p] = \mathcal{J}_{\mathcal{G}}^\varepsilon[u, \phi] + \langle \delta_\phi \mathcal{E}, p \rangle,$$

where \mathcal{G} stands for \mathcal{W} or \mathcal{C} , p denotes the Lagrange multiplier, and $(\phi - \text{id})|_{\Gamma_D} = p|_{\Gamma_D} = 0$. The associated necessary conditions are given by $0 = \delta_u L$ and $0 = \delta_\phi L = \delta_p L$ with $\delta_u L = \delta_u \mathcal{G} + \nu \delta_u \mathcal{V} + \eta \delta_u \mathcal{L}^\varepsilon + \delta_u \langle \delta_\phi \mathcal{E}, p \rangle$, $\delta_\phi L = \delta_\phi \mathcal{G} + \delta_\phi \langle \delta_\phi \mathcal{E}, p \rangle$, $\delta_p L = \delta_\phi \mathcal{E}$, and

$$\begin{aligned} \langle \delta_u \mathcal{V}, \vartheta \rangle &= \int_{\Omega} \frac{\partial \chi_{\mathcal{O}}[u]}{\partial u} \vartheta \, dx, \\ \langle \delta_u \mathcal{L}_{\text{MM}}^\varepsilon, \vartheta \rangle &= \int_{\Omega} \varepsilon \nabla u \cdot \nabla \vartheta + \frac{1}{2\varepsilon} \frac{\partial \Psi(u)}{\partial u} \vartheta \, dx, \\ \langle \delta_u \mathcal{W}, \vartheta \rangle &= \int_{\Omega} (1 - \delta) \frac{\partial \chi_{\mathcal{O}}[u]}{\partial u} \vartheta W(\mathcal{D}\phi) \, dx, \\ \langle \delta_\phi \mathcal{W}, \theta \rangle &= \int_{\Omega} ((1 - \delta)\chi_{\mathcal{O}}[u] + \delta)W_{,A}(\mathcal{D}\phi) : \mathcal{D}\theta \, dx, \\ \langle \delta_\phi \mathcal{C}, \theta \rangle &= \int_{\Gamma_N} F \cdot \theta \, da, \\ \langle \delta_u \langle \delta_\phi \mathcal{E}, p \rangle, \vartheta \rangle &= \int_{\Omega} (1 - \delta) \frac{\partial \chi_{\mathcal{O}}[u]}{\partial u} \vartheta W_{,A}(\mathcal{D}\phi) : \mathcal{D}p \, dx, \\ \langle \delta_\phi \langle \delta_\phi \mathcal{E}, p \rangle, \vartheta \rangle &= \int_{\Omega} ((1 - \delta)\chi_{\mathcal{O}}[u] + \delta)W_{,AA}(\mathcal{D}\phi) \mathcal{D}\theta : \mathcal{D}p \, dx \end{aligned}$$

for scalar and vector-valued test functions ϑ and θ , respectively, with $\theta|_{\Gamma_D} = 0$.

Furthermore, for a sufficiently smooth phase field u and deformation ϕ satisfying the equilibrium constraint, we may locally regard ϕ as a function $\phi[u]$. Then, by the adjoint method, the derivative of $\tilde{\mathcal{J}}_{\mathcal{G}}^\varepsilon[u] := \mathcal{J}_{\mathcal{G}}^\varepsilon[u, \phi[u]]$ with respect to u in direction ϑ is given as

$$\langle \delta_u \tilde{\mathcal{J}}_{\mathcal{G}}^\varepsilon, \vartheta \rangle = \langle \delta_u \mathcal{J}_{\mathcal{G}}^\varepsilon, \vartheta \rangle + \langle \delta_u \langle \delta_\phi \mathcal{E}, p \rangle, \vartheta \rangle,$$

where for fixed u , the deformation $\phi[u]$ and the Lagrange multiplier p solve $0 = \delta_\phi L = \delta_p L$ with the corresponding Dirichlet boundary conditions at Γ_D . This directional derivative can be used in gradient descent algorithms to find the optimal phase field u .

Concerning the discretisation, as in the previous chapters, we will approximate the phase field u and deformation ϕ by continuous, piecewise multilinear finite element functions U and Φ on a regular mesh on $\Omega = [0, 1]^d$ with $2^L + 1$ nodes in each space direction (see Section 4.3 for details). The different energy terms \mathcal{W} , \mathcal{V} , and $\mathcal{L}_{\text{MM}}^\varepsilon$ are approximated by third order Gaussian quadrature on each grid cell. In our applications, Γ_D and Γ_N are chosen as the union of several grid cell faces so that Γ_N , too, is discretised in the canonical way by a regular mesh on which a continuous, piecewise multilinear finite element approximation of the surface load F can be defined. \mathcal{C} is then also computed on that finite element mesh.

7.4.2 Inner minimisation to find equilibrium deformation

We aim at a gradient descent type algorithm for the (discretised) phase field U , where in each step we first minimise $\mathcal{E}[U, \Phi]$ to obtain a finite element approximation $\Phi[U]$ to the equilibrium deformation and then use this deformation to evaluate the energy $\mathcal{J}_{\mathcal{W}}^\varepsilon[U, \Phi[U]]$ or $\mathcal{J}_{\mathcal{C}}^\varepsilon[U, \Phi[U]]$ and its Gâteaux derivative with respect to U . The inner minimisation of $\mathcal{E}[U, \Phi]$ for Φ has to meet particularly strong requirements. First of all, the optimal deformation $\Phi[U]$ has to be accurately found in order to enable a correct evaluation of the objective energy and to obtain a good approximation to the Gâteaux derivative which can then be used to compute a descent direction. Second, since the minimisation has to be performed for each energy evaluation, we need a fast convergence. Finally, the optimisation method has to be very robust and should reliably lead to a (local) minimum.

The latter robustness requirement is particularly related to the use of a nonlinearly elastic energy: In the presence of buckling instabilities, there is typically an unstable or metastable (meaning that small perturbations suffice to abandon the state), non-buckled state of the deformation Φ which more or less corresponds to the deformation in the linearised elastic setting. This state is associated with a saddle point of the energy $\mathcal{E}[U, \Phi]$, which has to be robustly bypassed by the minimisation method. While simple gradient descent type methods tend to slow down considerably in the vicinity of such points, the basic Newton algorithm is prone to converging exactly against this saddle point. Hence, we will need a more sophisticated technique. Furthermore, the energy landscape in the nonlinear regime is typically characterised by long, deep, narrow and bent valleys. These valleys may be interpreted as the paths along which the material can be deformed, and leaving these valleys will rapidly lead to unphysical states such as local material interpenetration and thus the break-down of the minimisation.

The demand for high accuracy and fast convergence calls for a Newton-based minimisation method. Robustness can be ensured by an appropriate step size control, combined with a detection of saddle points. Trust region methods represent a very reliable minimisation technique that satisfies all the above issues. They are iterative methods that generate a descending sequence x_1, x_2, \dots for an objective functional $f : \mathbb{R}^N \rightarrow \mathbb{R}$, $x \mapsto f(x)$. At each step i , the objective functional is approximated by a quadratic

model m_i which is minimised inside a trust region around x_i to obtain a new guess x_{i+1} . If the decrease of the objective functional sufficiently agrees with the decrease of the quadratic model, the step is accepted and the trust region enlarged; otherwise, the trust region is shrunken.

The subtleties of a trust region method lie in its treatment of the so-called trust region subproblem to minimise the quadratic model within the trust region. In our computations, we chose to implement the algorithm proposed in [40, Algorithm 7.3.4]. At step i , the quadratic model m_i is given as the second order Taylor expansion of the energy $f(x)$ about x_i ,

$$m_i(x) = f(x_i) + \nabla f(x_i) \cdot (x - x_i) + \frac{1}{2}H(x_i)(x - x_i) \cdot (x - x_i),$$

where $H(x_i)$ shall denote the Hessian of f at x_i . To minimise this model within a circular trust region around x_i of radius Δ , the smallest positive scalar ξ is sought such that $H_i(\xi) := H(x_i) + \xi \text{id}$ becomes positive definite and the global minimum of

$$m_i^\xi(x) = f(x_i) + \nabla f(x_i) \cdot (x - x_i) + \frac{1}{2}H_i(\xi)(x - x_i) \cdot (x - x_i)$$

lies within the trust region. The positive definiteness of the quadratic operator is checked via a Cholesky factorisation $H_i(\xi) = LL^T$, which also serves to find the minimum of $m_i^\xi(x)$ by solving the linear system of equations $H_i(\xi)(x - x_i) = -\nabla f(x_i)$ that results from the optimality conditions. Additionally, the eigendirection belonging to the smallest eigenvalue of $H_i(\xi)$ is approximated by a technique which aims to find a vector v such that $L^{-1}v$ is large. This eigendirection is essentially employed to bypass saddle points. The scalar ξ is itself obtained by a Newton iteration which is safeguarded by a number of sophisticated bounds on ξ (see [40] for details). The Cholesky factorisation is performed using the CHOLMOD package from Davis et al. [48, 32], where a matrix reordering ensures a minimum fill-in.

In our setting, the minimisation variable x is the vector Φ of nodal values of the finite element deformation Φ , and the energy function f is given by $\mathcal{E}[U, \Phi]$ for a fixed U . The gradient ∇f and the Hessian matrix H can be represented as

$$\begin{aligned} \nabla f &= (\langle \delta_\phi \mathcal{E}, \varphi_i e_j \rangle)_{(i,j) \in \hat{I}_h \times \{1, \dots, d\}} = \\ &\left(\int_{\Omega} ((1 - \delta)\chi_{\mathcal{O}}[U] + \delta) W_{,A}(\mathcal{D}\Phi) : \mathcal{D}(\varphi_i e_j) \, dx - \int_{\Gamma_N} F \cdot (\varphi_i e_j) \, da \right)_{(i,j) \in \hat{I}_h \times \{1, \dots, d\}} \end{aligned}$$

and

$$\begin{aligned} H &= (\langle \delta_\phi \mathcal{E}, \varphi_i e_j, \varphi_k e_l \rangle)_{(i,j),(k,l) \in \hat{I}_h \times \{1, \dots, d\}} = \\ &\left(\int_{\Omega} ((1 - \delta)\chi_{\mathcal{O}}[U] + \delta) \langle W_{,AA}(\mathcal{D}\Phi), \mathcal{D}(\varphi_i e_j), \mathcal{D}(\varphi_k e_l) \rangle \, dx \right)_{(i,j),(k,l) \in \hat{I}_h \times \{1, \dots, d\}} \end{aligned}$$

for the set \hat{I}_h of node indices in $\Omega \setminus \Gamma_D$, the finite element basis functions φ_i , $i \in \hat{I}_h$, and the canonical Euclidean basis e_1, \dots, e_d .

7.4.3 Optimisation for the phase field

Concerning the outer optimisation for U , we apply a Davidon–Fletcher–Powell quasi-Newton method, which—again expressed for the minimisation of a function $f : \mathbb{R}^N \rightarrow \mathbb{R}$, $x \mapsto f(x)$ —uses the update formula

$$B_{k+1} = B_k + \frac{\Delta x_k \Delta x_k^T}{g_k^T \Delta x_k} - \frac{B_k g_k g_k^T B_k^T}{g_k^T B_k g_k}$$

to approximate the inverse of the Hessian of f in the $(k+1)$ th step using the latest update $\Delta x_k = x_{k+1} - x_k$ and the difference $g_k = \nabla f(x_{k+1}) - \nabla f(x_k)$ between the gradients. The descent direction p_k is then chosen as $-B_k \nabla f(x_k)$, and the step length τ_k is determined to satisfy the strong Wolfe conditions,

$$\begin{aligned} f(x_k + \tau_k p_k) &\leq f(x_k) + c_1 \tau_k \nabla f(x_k) \cdot p_k, \\ |\nabla f(x_k + \tau_k p_k) \cdot p_k| &\leq c_2 |\nabla f(x_k) \cdot p_k| \end{aligned}$$

for $c_1 = 0.5$, $c_2 = 0.9$. Furthermore, we reset B_k to the identity every tenth step to restrict memory usage and to ensure a descent at least as good as gradient descent.

The gradient of the objective functional $\mathcal{J}_{\mathcal{G}}^\varepsilon[U, \Phi[U]]$ (for $\mathcal{G} = \mathcal{W}$ or $\mathcal{G} = \mathcal{C}$) with respect to U is here computed via the adjoint method as described in Section 7.4.1. We first solve

$$\langle \delta_\phi \langle \delta_\phi \mathcal{E}, \Psi \rangle, P \rangle = -\langle \delta_\phi \mathcal{J}_{\mathcal{G}}^\varepsilon, \Psi \rangle$$

for the finite element Lagrange multiplier P under the constraint $P|_{\Gamma_D} = 0$, where Ψ runs over all vector-valued finite element functions that are zero on Γ_D . In terms of finite element operators, this can be expressed as the linear system $H\mathbf{P} = \mathbf{R}$, where \mathbf{P} denotes the vector of nodal values of P on $\Omega \setminus \Gamma_D$, the matrix H has been given above, and the right-hand side reads

$$\begin{aligned} \mathbf{R} &= \left(\int_{\Omega} ((1 - \delta)\chi_{\mathcal{O}}[U] + \delta)W_{,A}(\mathcal{D}\Phi[U]) : \mathcal{D}(\varphi_i e_j) \, dx \right)_{(i,j) \in \hat{I}_h \times \{1, \dots, d\}} \\ \text{or } \mathbf{R} &= \left(\int_{\Gamma_N} F \cdot (\varphi_i e_j) \, da \right)_{(i,j) \in \hat{I}_h \times \{1, \dots, d\}}, \end{aligned}$$

depending on whether $\mathcal{J}_{\mathcal{W}}^\varepsilon[U, \Phi[U]]$ or $\mathcal{J}_{\mathcal{C}}^\varepsilon[U, \Phi[U]]$ is minimised. Then we obtain the gradient of the objective functional with respect to U as

$$\langle \delta_u \mathcal{J}_{\mathcal{G}}^\varepsilon, \varphi_i \rangle + \langle \delta_u \langle \delta_\phi \mathcal{E}, P \rangle, \varphi_i \rangle$$

for all $i \in \tilde{I}_h$, where \tilde{I}_h represents the set of node indices in Ω at which U is not fixed by a Dirichlet condition, and where the expressions for the Gâteaux derivatives are provided in Section 7.4.1.

7.4.4 Embedding the optimisation in a multiscale approach

In order to enhance convergence and to avoid local minima, we pursue a multiscale approach once more, using the same hierarchy of dyadic grid resolutions and prolongation techniques as described in Section 4.3. We first perform the minimisation for a coarse spatial discretisation and then successively prolongate and refine the result on finer grids. The phase field scale parameter ε is coupled to the grid size h via $\varepsilon = h$ in order to allow a sufficient resolution of the interface. Finally, it is sometimes advantageous to take a smaller value of ν on coarse grids in order not to penalise the value $U = 1$ so strongly that intermediate values of U between -1 and 1 are preferred. As the grid gets finer, ν can be increased since the smaller value of ε forces the phase field values towards the pure phases -1 and 1 .

A brief overview over the entire algorithm in pseudo code notation reads as follows (bold capital letters represent vectors of nodal values, and \mathcal{G} stands for \mathcal{W} or \mathcal{C}):

```

EnergyRelaxation ( $F$ ) {
  initialise  $\Phi = \text{id}$  and  $\mathbf{U}_i = 0$ ,  $i \in \tilde{I}_h$ , on grid level  $l_0$ ;
  for grid level  $l = l_0$  to  $L$  {
    do {
       $\mathbf{U}^{\text{old}} = \mathbf{U}$ ;
      minimise  $\mathcal{E}[U, \Phi]$  for  $\Phi$  by a trust region method to obtain  $\Phi[U]$ ;
      evaluate  $\mathcal{J}_{\mathcal{G}}^{\varepsilon}[U, \Phi[U]]$ ;
      compute the dual variable  $P$  by solving the linear system
         $\langle \delta_{\phi} \langle \delta_{\phi} \mathcal{E}, \Psi \rangle, P \rangle = -\langle \delta_{\phi} \mathcal{J}_{\mathcal{G}}^{\varepsilon}, \Psi \rangle \quad \forall \Psi$ ;
      compute the derivative of  $\tilde{\mathcal{J}}_{\mathcal{G}}^{\varepsilon}[U] := \mathcal{J}_{\mathcal{G}}^{\varepsilon}[U, \Phi[U]]$  with respect to  $\mathbf{U}$  as
         $\mathbf{V} := (\langle \delta_u \mathcal{J}_{\mathcal{G}}^{\varepsilon}, \varphi_i \rangle + \langle \delta_u \langle \delta_{\phi} \mathcal{E}, P \rangle, \varphi_i \rangle)_{i \in \tilde{I}_h}$ ;
      compute an approximate inverse Hessian  $B$  by the DFP method;
      compute a descent direction  $\mathbf{D} := -B\mathbf{V}$ ;
      perform a descent step
         $\mathbf{U} = \mathbf{U}^{\text{old}} - \tau \mathbf{D}$ 
        with Wolfe step size control for  $\tau$ ;
    } while ( $|\mathbf{U}^{\text{old}} - \mathbf{U}| \geq \text{Threshold}$ );
    if ( $l < L$ ) prolongate  $U, \Phi$  onto the next grid level;
  }
}

```

7.5 Experiments

The effect of using nonlinear instead of linearised elasticity has already been explored in Figure 7.2, where we compare optimal cantilever shapes for loads of different magnitudes, using $\mathcal{J}_{\mathcal{C}}^{\varepsilon}$ as objective functional. White and black regions correspond to the phases $u = 1$ and $u = -1$, respectively. Of course, increasing the load results in stronger deformations

and thus higher values of \mathcal{C} (and \mathcal{W}) so that the shapes would naturally become thicker and more strutted in order to balance the compliance with the volume costs \mathcal{V} and the regularisation $\mathcal{L}_{\text{MM}}^\varepsilon$. To make the optimal designs comparable and to reveal the pure influence of introducing geometric and material nonlinearity, the weights ν and η of \mathcal{V} and $\mathcal{L}_{\text{MM}}^\varepsilon$ have to be increased in parallel. Since the compliance scales quadratically with the load (at least for small deformations in the regime of linearised elasticity), ν and η are chosen such that $\frac{F^2}{\nu}$ and $\frac{F^2}{\eta}$ stay constant.

As discussed in Section 7.2, compared to optimal designs with linearised elasticity, the symmetry of the cantilever design is broken due to the nonlinear influence of the loading direction. We observe that as the load increases, a structure evolves which exhibits a single thick beam at the top that is supported from below by several thinner struts extending from the wall to the point where the load is applied as well as to two or three other points along the beam. These struts are themselves suspended from thread-like structures.

In the beginning, we have stated and discussed the different possibilities to extend the notion of compliance to the setting of nonlinear elasticity. In particular we have considered the change of external potential \mathcal{C} and the internally stored elastic energy \mathcal{W} . The compliance minimisation yields different results depending on our choice: While the use of \mathcal{C} will produce rather rigid constructions that allow only small displacements, the use of \mathcal{W} does in principle allow for large deformations as long as the final equilibrium state is not heavily strained. A third possibility would be to employ the difference, $-\mathcal{E} = \mathcal{C} - \mathcal{W}$, which describes the energy dissipation during the deformation to the equilibrium state and equals \mathcal{W} in linearised elasticity.

A comparison of the three possibilities reveals a complex interplay between the volume costs and the mechanical energy (Figure 7.10). We first observe that the obtained shapes are plausible in the sense that each of them indeed has the minimum value of its associated objective function ($\mathcal{J}_{\mathcal{C}}^\varepsilon$, $\mathcal{J}_{\mathcal{W}}^\varepsilon$, or $\mathcal{J}_{-\mathcal{E}}^\varepsilon$, respectively) among all three. However, we also notice that the various energy contributions do not differ significantly between the three designs. Nevertheless, they do look quite different: While the minimiser of $\mathcal{J}_{-\mathcal{E}}$ is almost symmetric, the minimiser of $\mathcal{J}_{\mathcal{W}}$ is strongly asymmetric. In particular, the bottommost beam, which will be compressed during the deformation, becomes thinner from $\mathcal{J}_{-\mathcal{E}}$ to $\mathcal{J}_{\mathcal{C}}$ to $\mathcal{J}_{\mathcal{W}}$, thereby allowing stronger displacements. This agrees with the simplified model example in Section 7.1. The internal energy \mathcal{W} seems to be not very sensitive to the thickness of the bottommost beam, while the potential change \mathcal{C} strongly is. Therefore, it pays off to save volume at the cost of a slightly increasing \mathcal{W} in order to minimise $\mathcal{J}_{\mathcal{W}}$.

In fact, the model example in Section 7.1 already suggests that the middle design in Figure 7.10 is only locally optimal. Indeed, if we initialise the shape as a simple rod similar to Figure 7.1, then we obtain the design shown in Figure 7.11 with a much lower objective function value.

Apart from the employed hyperelastic material law, there are two weighting parameters that need to be tuned in order to obtain sensible results. Consequently, we have to

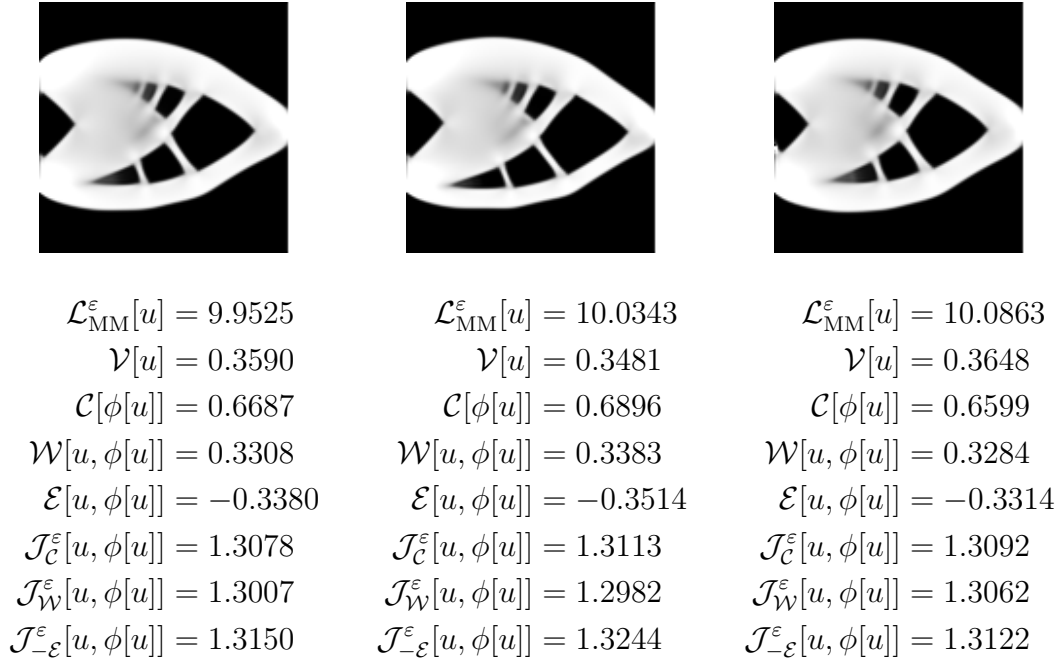


Figure 7.10: Optimal cantilever design for minimising $\mathcal{J}_{\mathcal{C}}^\epsilon$, $\mathcal{J}_{\mathcal{W}}^\epsilon$, $\mathcal{J}_{-\mathcal{E}}^\epsilon$ (from left to right), taking the same parameters as in Figure 7.2 for the case $\hat{F} = 4$ (in these computations, ϵ was chosen half the grid size h in order to obtain acceptable phase fields already for a resolution of 129×129). Minimisation of $\mathcal{J}_{-\mathcal{E}}^\epsilon$ yields the most symmetric shape, while minimisation of $\mathcal{J}_{\mathcal{W}}^\epsilon$ yields the most asymmetric one with the compressed struts being much thinner than in the other cases.

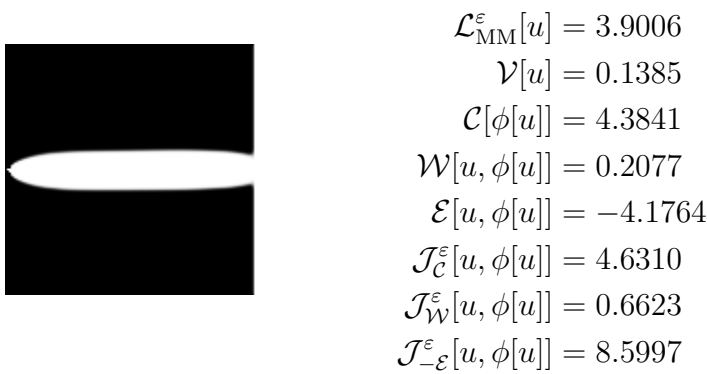


Figure 7.11: Initialising the phase field as a single rod, a minimisation of $\mathcal{J}_{\mathcal{W}}^\epsilon$ retrieves the above optimal cantilever design with a much lower objective function value than the design in Figure 7.10, middle.

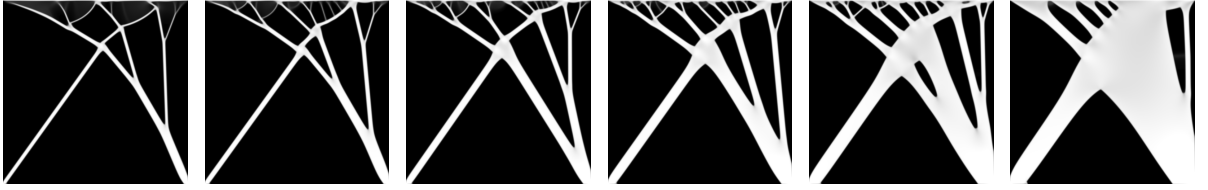


Figure 7.12: Optimal designs for the same load case as in Figure 7.6 with parameters $F = 2^4 \cdot 10^{-3} \cdot \sqrt{2}$, $\lambda = \mu = 80$, $\eta = 2^k \cdot 10^{-8}$, $\nu = 2^{5+k} \cdot 10^{-7}$ for $k = 6, \dots, 1$ from left to right (resolution 512×512). The load is chosen very small so as to stay in the regime of linearised elasticity such that buckling cannot occur for the fine shapes on the left.

study the influence of the different energy contributions. The impact of volume penalisation and perimeter regularisation seems clear; the former prefers thinner structures while the latter tends to reduce the degree of cross-linking. The question arises how the compliance term affects the optimal shape, whether by thickening its single components or by producing more strutted shapes. Experiments show that both mechanisms occur until a point at which the different material parts grow so wide that they start to merge and eliminate any fine structure (Figure 7.12).

To explore the effect of varying the two weight parameters ν and η systematically, let us sample the two parameter family of (at least locally) optimal shapes which is generated by η and ν . Before doing so, note that generally, the maximally reachable resolution of our optimal designs is restricted due to computation time, which in turn also fixes the smallest resolvable scale parameter ε . Since the phase field u is only forced towards the pure phases $u \in \{-1, 1\}$ for the limit $\varepsilon \rightarrow 0$, there will inherently always remain intermediate values (which can also be observed in some of the previous examples). One particular reason for this lies in the strong interaction between the non-convex chemical potential $\eta \frac{9}{32} \frac{1}{\varepsilon} (u^2 - 1)^2$ with the two energy minimising phases $u \in \{-1, 1\}$ and the convex volume costs $\frac{\nu}{4} (u + 1)^2$. For too large ε , that is, for $\varepsilon \geq \frac{\eta}{\nu} \frac{9}{4}$, their sum becomes convex so that there are no longer two preferred phases! Due to the significance of $\frac{\nu}{\eta}$ in this lower bound, in our parameter study we shall choose to vary $\frac{\nu}{\eta}$ and η instead of ν and η .

Figure 7.13 shows cantilever designs which were obtained by choosing the middle computation as the reference setting and then doubling or halving $\frac{\nu}{\eta}$ as well as η . Apparently, for fixed ε there is a regime of $\frac{\nu}{\eta}$ for which reasonable shapes are obtained, corresponding to the middle column. For larger values, phase field densities between -1 and 1 are not sufficiently penalised, and for smaller values we obtain bulky designs without any fine structure. The effect of varying η for a constant volume weight ν can be seen along the descending diagonals; these designs indeed seem to possess similar volumes, but a distinct amount of fine structure. Along the horizontal direction, just ν changes, and the design volumes can be seen to decrease to the right.

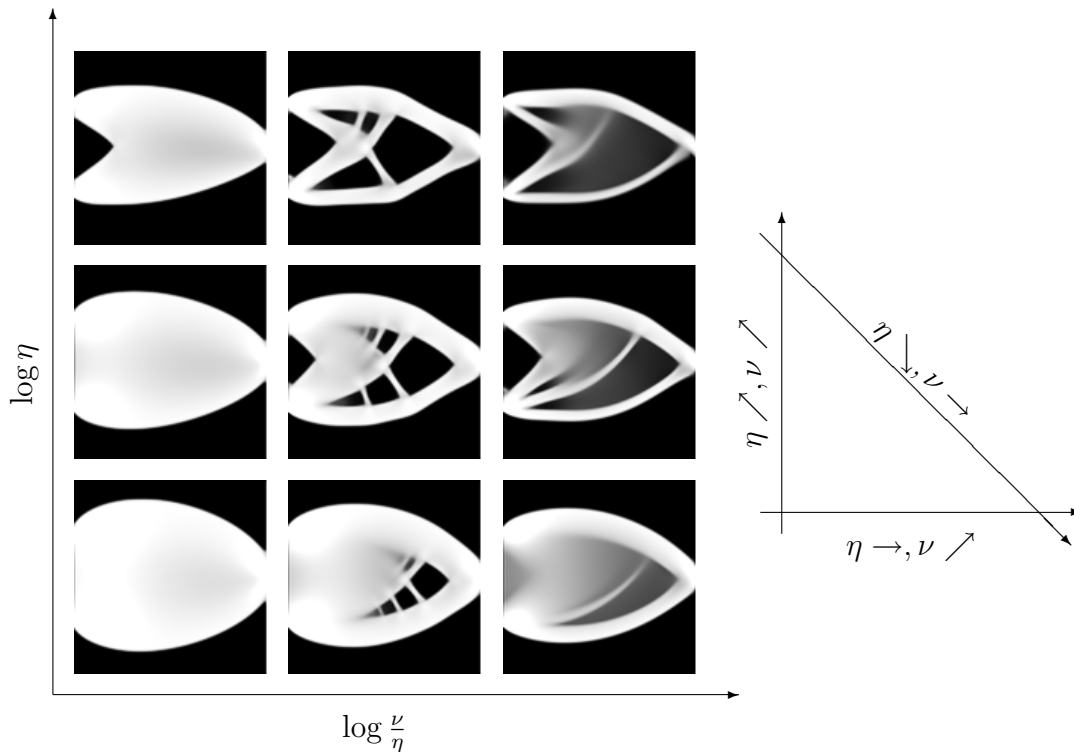


Figure 7.13: Optimal cantilever designs for different parameter values. The middle design is the same as in Figure 7.10, left; the top and bottom row as well as the right and left column are obtained by doubling and halving η and $\frac{\nu}{\eta}$, respectively.

Naturally, the energy landscape associated with $\mathcal{J}_C^\varepsilon$, $\mathcal{J}_W^\varepsilon$, and $\mathcal{J}_{-\varepsilon}^\varepsilon$ is quite complicated and allows for multiple local minima. In order to reduce the influence of initialisation and to obtain satisfactory results with a sufficiently low objective function value, in the above examples we pursued a multiscale approach with an initial optimisation on a coarse resolution (to find a good large scale structure) and successive refinement. For initialisation on the coarsest level, each pixel of the phase field is taken randomly from a uniform distribution on $[-0.1, 0.1]$. Also, we started with a small value of ν on the coarse grid and then doubled it each prolongation until its final value on the finest grid level. In this way we maintain a constant ratio $\frac{\eta}{\varepsilon\nu}$ over all grid levels so that volume costs and chemical potential already balance each other on the coarser grid levels and the coarse phase fields are already quite close to an optimal design (compare Figures 7.14 to 7.18).

Alternatively, one could employ the same value of ν on all grid levels or start the optimisation directly on the finest level without a multiscale approach. While all three approaches yield the same result for many parameter constellations (which is not too surprising especially for the rightmost column in Figure 7.13, for example, where we have reached the regime in which the sum of volume costs and chemical potential are convex),

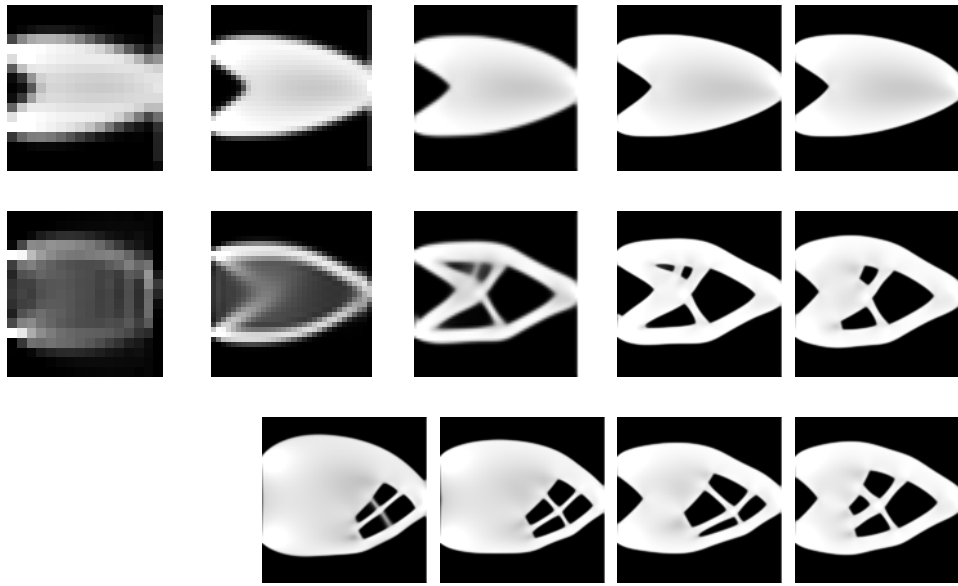


Figure 7.14: Intermediate results of the optimisation for the top left case of Figure 7.13. The first three images in the top row show the results of the multiscale algorithm on grid levels four to six, where ν has been doubled at each prolongation. The rightmost two images represent the shape after 300 quasi-Newton iterations on grid level seven as well as the final result. The middle row shows the same results, only that this time ν is the same on all grid levels. In the bottom row, the optimisation was started directly on grid level seven without a multiscale approach. The intermediate shapes after 300, 600, and 3000 iterations as well as the final result are depicted. The final values of the objective function $\mathcal{J}_C^\varepsilon$ are 1.3500, 1.3229, 1.3239, respectively.

this does not hold for the cases depicted in Figures 7.14 to 7.18, where the progress of the optimisation algorithm is shown. In general, however, the achieved objective function values are very close to each other, and in some cases it seems that at least two of the tree approaches would eventually converge against the same shape if they were not terminated due to too small progress. Note that for the cases in Figures 7.14 and 7.15, some approaches seem not to be able to create holes inside the bulky shape.

As a final example, we have computed an optimal design for a bridge-like structure with two pointwise Dirichlet boundary conditions and a uniform downward surface load (Figure 7.19). Apparently, the optimal design is to suspend the bottom edge from an arch extending between both fixing points. In order to reduce computation time for this example (especially during the Cholesky-factorisation), we did not choose the entire unit square $[0, 1]^2$ as the computational domain, but instead updated the computational domain every 100 iterations as the region $\{x \in [0, 1]^2 : u(x) > -0.95\}$, dilated by three pixels. This update does not hamper convergence, and the final, actual computational

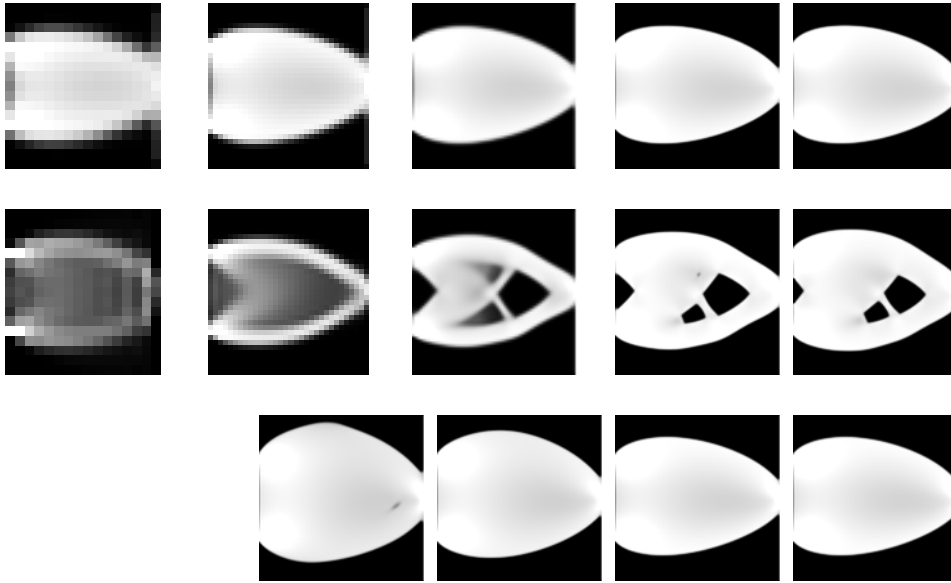


Figure 7.15: Intermediate results of the optimisation for the middle left case of Figure 7.13. The rows correspond to three different variants of the algorithm as explained in Figure 7.14. The final values of the objective function $\mathcal{J}_C^\varepsilon$ are 0.9192, 0.9160, 0.9192, respectively.

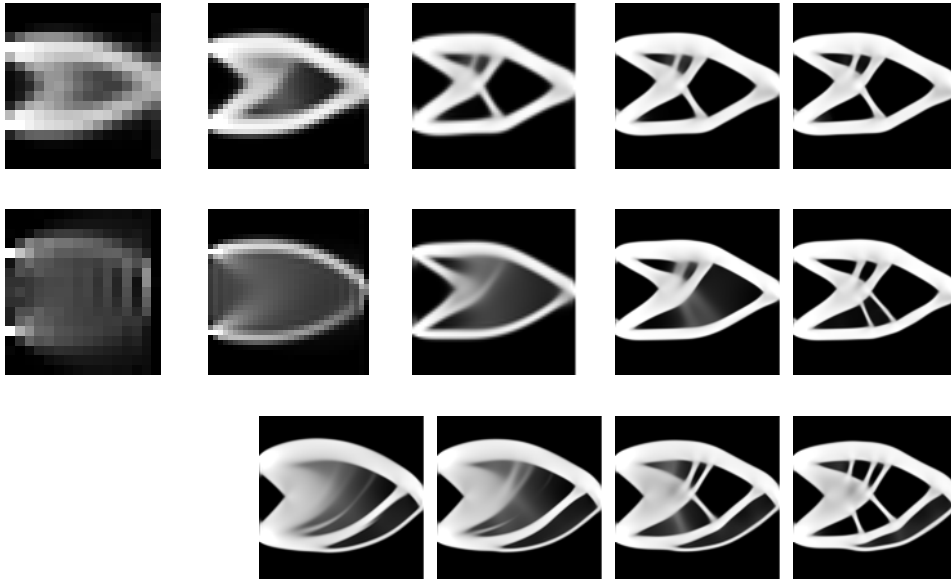


Figure 7.16: Intermediate results of the optimisation for the top middle case of Figure 7.13. The rows correspond to three different variants of the algorithm as explained in Figure 7.14. The final values of the objective function $\mathcal{J}_C^\varepsilon$ are 1.9022, 1.9617, 1.8967, respectively.

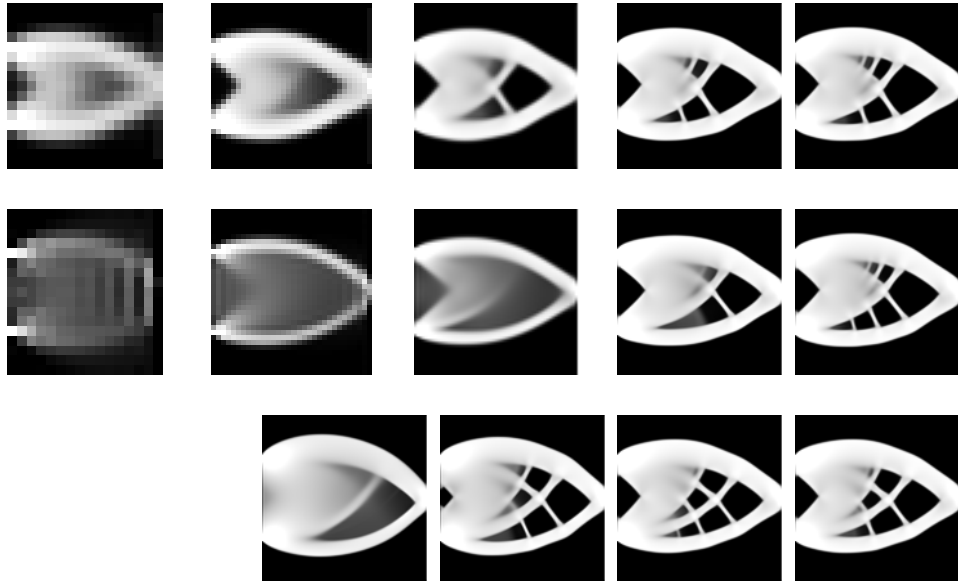


Figure 7.17: Intermediate results of the optimisation for the middle case of Figure 7.13. The rows correspond to three different variants of the algorithm as explained in Figure 7.14. The final values of the objective function $\mathcal{J}_C^\varepsilon$ are 1.2727, 1.2728, 1.2729, respectively.

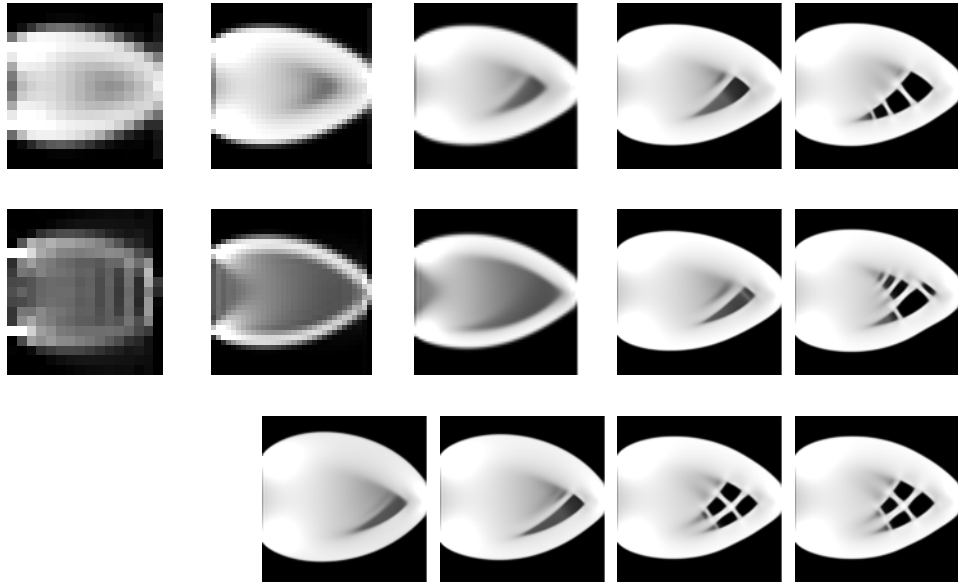


Figure 7.18: Intermediate results of the optimisation for the bottom middle case of Figure 7.13. The rows correspond to three different variants of the algorithm as explained in Figure 7.14. The final values of the objective function $\mathcal{J}_C^\varepsilon$ are 0.8870, 0.8872, 0.8872, respectively.

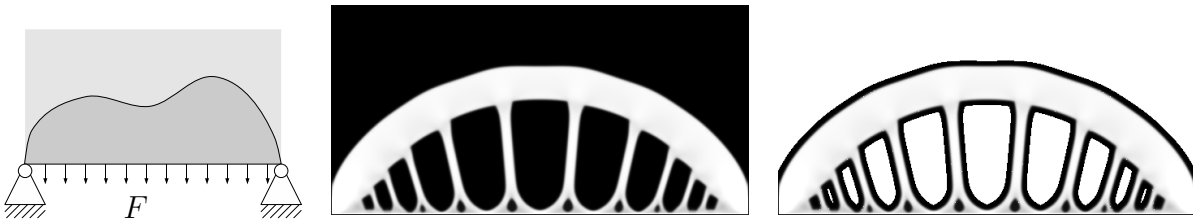


Figure 7.19: Sketch of the design problem and optimal design for $F = 1.2$, $\lambda = \mu = 80$, $\eta = 10^{-4}$, $\nu = 2^5 \cdot 10^{-3}$. The right image shows the final, actual computational domain.

domain is shown in Figure 7.19, right.

Acknowledgements

I am especially indebted to Martin Rumpf for the supervision and support of my thesis. Furthermore I would like to thank Leah Bar, Patrick Penzler, and Guillermo Sapiro for the excellent cooperation on projects within the thesis. I am also deeply grateful to Benjamin Berkels, Martin Lenz, and Ole Schwen for their assistance with any hardware, software, and programming problems. Also, I thank them as well as Stefan von Deylen, Martina Teusner, and all members of Martin Rumpf's group at the Institute for Numerical Simulation, Bonn, for our numerous valuable mathematical discussions.

For financial as well as non-material support I thank the Bonn International Graduate School in Mathematics, the Hausdorff Center for Mathematics at Bonn University, and the Studienstiftung des deutschen Volkes.

Thanks also to Carlo Schaller and Marc Kotowski, Hôpitaux Universitaire de Genève, Switzerland, for the provision of cerebral CT scans and discussions on medical imaging, to Werner Bautz, radiology department at the university hospital Erlangen, Germany, for providing CT data of kidneys, to Heiko Schlarb from Adidas, Herzogenaurach, Germany, for providing 3D scans of feet, and to Bruno Wirth, urology department at the Hospital zum Hl. Geist, Kempen, Germany, for providing thorax CT scans.

Last but not least I owe much to my family and my girlfriend for their continuous support.



Bibliography

- [1] Kazuhisa Abe, Shunsuke Kazama, and Kazuhiro Koro. A boundary element approach for topology optimization problem using the level set method. *Communications in Numerical Methods in Engineering*, 23:405–416, 2007.
- [2] G. Allaire. Topology optimization with the homogenization and the level-set method. pages 1–13. Kluwer Academic Pub, November 2004.
- [3] G. Allaire, E. Bonnetier, G. Francfort, and F. Jouve. Shape optimization by the homogenization method. *Numerische Mathematik*, 76:27–68, 1997.
- [4] G. Allaire, F. Jouve, and H. Maillot. Topology optimization for minimum stress design with the homogenization method. *Struct Multidisc Optim*, 28:87–98, 2004.
- [5] Grégoire Allaire, Francois Jouve, and Anca-Maria Toader. A level-set method for shape optimization. *C. R. Acad. Sci. Paris, Série I*, 334:1125–1130, 2002.
- [6] Grégoire Allaire, Francois Jouve, and Anca-Maria Toader. Structural optimization using sensitivity analysis and a level-set method. *Journal of computational physics*, 194:363–393, 2004.
- [7] L. Ambrosio and V. M. Tortorelli. Approximation of functionals depending on jumps by elliptic functionals via Γ -convergence. *Comm. Pure Appl. Math.*, 43:999–1036, 1990.
- [8] L. Ambrosio and V. M. Tortorelli. On the approximation of free discontinuity problems. *Bollettino dell’Unione Matematica Italiana, Sezione B*, 6(7):105–123, 1992.
- [9] Luigi Ambrosio and Giuseppe Buttazzo. An optimal design problem with perimeter penalization. *Calculus of Variations and Partial Differential Equations*, 1:55–69, 1993.
- [10] Rubén Ansola, Estrella Veguería, Javier Canales, and José A. Tárrago. A simple evolutionary topology optimization procedure for compliant mechanism design. *Finite Elements in Analysis and Design*, 44(1–2):53–62, 2007.
- [11] J. M. Ball. Convexity conditions and existence theorems in nonlinear elasticity. *Archive of Rational Mechanics and Analysis*, 63:337–403, 1977.

- [12] J.M. Ball. Global invertibility of Sobolev functions and the interpenetration of matter. *Proc. Roy. Soc. Edinburgh*, 88A:315–328, 1981.
- [13] M. Faisal Beg, Michael I. Miller, Alain Trouvé, and Laurent Younes. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *International Journal of Computer Vision*, 61(2):139–157, February 2005.
- [14] K. K. Bhatia, J. V. Hajnal, A. Hammers, and D. Rueckert. Similarity metrics for groupwise non-rigid registration. In N. Ayache, S. Ourselin, and A. Maeder, editors, *Medical Image Computing and Computer-Assisted Intervention, MICCAI 2007*, volume 4792 of *LNCS*, pages 544–552, 2007.
- [15] K. K. Bhatia, J. V. Hajnal, B. K. Puri, A. D. Edwards, and D. Rueckert. Consistent groupwise non-rigid registration for atlas construction. In *IEEE International Symposium on Biomedical Imaging: Nano to Macro*, volume 1, pages 908–911, 2004.
- [16] M. J. Black and P. Anandan. A framework for the robust estimation of optical flow. In *Fourth International Conference on Computer Vision, ICCV-93*, pages 231–236, 1993.
- [17] B. Bourdin and A. Chambolle. Design-dependent loads in topology optimization. *ESAIM Control Optim. Calc. Var.*, 9:19–48, 2003.
- [18] Andrea Braides. Γ -convergence for beginners, volume 22 of *Oxford Lecture Series in Mathematics and its Applications*. Oxford University Press, Oxford, 2002.
- [19] M. Bro-Nielsen and C. Gramkow. Fast fluid registration of medical images. In K. H. Höhne and R. Kikinis, editors, *Visualization in Biomedical Computing: 4th International Conference, VBC*, volume 1131 of *LNCS*, pages 267–276, 1996.
- [20] A. Bronstein, M. Bronstein, and R. Kimmel. *Numerical Geometry of Non-Rigid Shapes*. Monographs in Computer Science. Springer, 2008.
- [21] M. Burger. A framework for the construction of level set methods for shape optimization and reconstruction. *Interfaces and Free Boundaries*, 5:301–329, 2003.
- [22] Martin Burger and Roman Stainko. Phase-field relaxation of topology optimization with local stress constraints. *SIAM Journal on Control and Optimization*, 45(4):1447–1466, 2006.
- [23] Bernard Chalmond and Stéphane C. Girard. Nonlinear modeling of scattered multivariate data and its application to shape change. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(5):422–432, 1999.

-
- [24] Antonin Chambolle. A density result in two-dimensional linearized elasticity, and applications. *Archive for Rational Mechanics and Analysis*, 167(3):211–233, 2003.
- [25] T. Chan and L. Vese. An active contour model without edges. In M. Nielsen, P. Johansen, O. F. Olsen, and J. Weickert, editors, *Scale-Space Theories in Computer Vision. Second International Conference, Scale-Space '99, Corfu, Greece, September 1999*, Lecture Notes in Computer Science; 1682, pages 141–151. Springer, 1999.
- [26] Tony F. Chan and Luminita A. Vese. A level set algorithm for minimizing the Mumford-Shah functional in image processing. Technical Report 00-13, University of California, Los Angeles, April 2000.
- [27] Tony F. Chan and Luminita A. Vese. Active contours without edges. *IEEE Transactions on Image Processing*, 10(2):266–277, 2001.
- [28] Tony F. Chan and Luminita A. Vese. A level set algorithm for minimizing the Mumford-Shah functional in image processing. In *IEEE/Computer Society Proceedings of the 1st IEEE Workshop on Variational and Level Set Methods in Computer Vision*, pages 161–168, 2001.
- [29] G. Charpiat, O. Faugeras, R. Keriven, and P. Maurel. Distance-based shape statistics. In *Acoustics, Speech and Signal Processing, 2006 (ICASSP 2006)*, volume 5, 2006.
- [30] G. Charpiat, P. Maurel, J.-P. Pons, R. Keriven, and O. Faugeras. Generalized gradients: Priors on minimization flows. *International Journal of Computer Vision*, 73(3):325–344, 2007.
- [31] Guillaume Charpiat, Olivier Faugeras, and Renaud Keriven. Approximations of shape metrics and application to shape warping and empirical shape statistics. *Foundations of Computational Mathematics*, 5(1):1–58, 2005.
- [32] Y. Chen, T. A. Davis, W. W. Hager, and S. Rajamanickam. Algorithm 887: CHOLMOD, supernodal sparse Cholesky factorization and update/downdate. *ACM Transactions on Mathematical Software*, 35(3):22:1–22:14, 2009.
- [33] M. Chipot and L. C. Evans. Linearization at infinity and lipschitz estimates in the calculus of variations. *Proceedings of the Royal Society of Edinburgh A*, 102(3–4):291–303, 1986.
- [34] G. Christensen, Richard D Rabbitt, and Michael I. Miller. Deformable templates using large deformation kinematics. *IEEE Transactions on Image Processing*, 5(10):1435–1447, October 1996.
- [35] G. E. Christensen, S. C. Joshi, and M. I. Miller. Volumetric transformations of brain anatomy. *IEEE Trans. Medical Imaging*, 16, no. 6:864–877, 1997.

- [36] G. E. Christensen, R. D. Rabbitt, and M. I. Miller. A deformable neuroanatomy textbook based on viscous fluid mechanics. In J. Prince and T. Runolfsson, editors, *Proc. 27th Annual Conf. Information Sci. and Systems*, pages 211–216, 1993.
- [37] G. E. Christensen, R. D. Rabbitt, and M. I. Miller. 3D brain mapping using a deformable neuroanatomy. *Phys. Med. Biol.*, 39(3):609–618, 1994.
- [38] P. G. Ciarlet. *Three-dimensional elasticity*. Elsevier Science Publishers B. V., 1988.
- [39] U. Clarenz, M. Droske, and M. Rumpf. Towards fast non-rigid registration. In *Inverse Problems, Image Analysis and Medical Imaging, AMS Special Session Interaction of Inverse Problems and Image Analysis*, volume 313, pages 67–84. AMS, 2002.
- [40] A. R. Conn, N. I. M Gould, and P. L. Toint. *Trust-Region Methods*. SIAM, 2000.
- [41] S. Conti, H. Held, M. Pach, M. Rumpf, and R. Schultz. Shape optimization under uncertainty - a stochastic programming perspective. *SIAM Journal on Optimization*, 19(4):1610–1632, 2008.
- [42] Sergio Conti, Harald Held, Martin Pach, Martin Rumpf, and Rüdiger Schultz. Risk averse shape optimization. *Siam Journal on Control and Optimization*, 2009. submitted.
- [43] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models—their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995.
- [44] Daniel Cremers, Timo Kohlberger, and Christoph Schnörr. Shape statistics in kernel space for variational image segmentation. *Pattern Recognition*, 36:1929–1943, 2003.
- [45] B. Dacorogna. *Direct methods in the calculus of variations*. Springer-Verlag, New York, 1989.
- [46] G. Dal Maso, J.M. Morel, and S. Solimini. A variational method in image segmentation: existence and approximation results. *Acta Math.*, 168(1-2):89–151, 1992.
- [47] Samuel Dambreville, Yogesh Rathi, and Allen Tannenbaum. A shape-based approach to robust image segmentation. In A. Campilho and M. Kamel, editors, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 4141 of *LNCS*, pages 173–183, 2006.

- [48] T. A. Davis and W. W. Hager. Dynamic supernodes in sparse Cholesky update/downdate and triangular solves. *ACM Transactions on Mathematical Software*, 35(4):27:1–27:23, 2009.
- [49] E. De Giorgi, M. Carriero, and A. Leaci. Existence theorem for a minimum problem with free discontinuity set. *Arch. Rat. Mech. and Anal.*, 108:195–218, 1989.
- [50] M. Droske and M. Rumpf. A variational approach to non-rigid morphological registration. *SIAM Journal on Applied Mathematics*, 64(2):668–687, 2004.
- [51] M. Droske and M. Rumpf. Multi scale joint segmentation and registration of image morphology. *IEEE Transaction on Pattern Recognition and Machine Intelligence*, 29(12):2181–2194, 2007.
- [52] D. Dupuis, U. Grenander, and M.I. Miller. Variational problems on flows of diffeomorphisms for image matching. *Quarterly of Applied Mathematics*, 56:587–600, 1998.
- [53] I. Eckstein, J.-P. Pons, Y. Tong, C.-C. Kuo, and M. Desbrun. Generalized surface flows for mesh processing. In *Eurographics Symposium on Geometry Processing*, 2007.
- [54] O. Faugeras and G. Hermosillo. Well-posedness of eight problems of multi-modal statistical image matching. Technical Report 4235, INRIA Sophia Antipolis, 2004.
- [55] O. Faugeras and G. Hermosillo. Well-posedness of two nonrigid multimodal image registration methods. *SIAM J. Appl. Math.*, 64:1550–1587, 2004.
- [56] Olivier Faugeras, Geoffray Adde, Guillaume Charpiat, Christophe Chéd'Hotel, Maureen Clerc, Thomas Deneux, Rachid Deriche, Gerardo Hermosillo, Renaud Keriven, Pierre Kornprobst, Jan Kybic, Christophe Lenglet, Lucero Lopez-Perez, Théo Papadopoulos, Jean-Philippe Pons, Florent Segonne, Bertrand Thirion, David Tschumperlé, Thierry Viéville, and Nicolas Wotawa. Variational, geometric, and statistical methods for modeling brain anatomy and function. *NeuroImage*, 23:S46–S55, 2004.
- [57] P. Thomas Fletcher, Conglin Lu, and Sarang Joshi. Statistics of shape via principal geodesic analysis on Lie groups. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR*, volume 1, pages 95–101, 2003.
- [58] P.T. Fletcher, Conglin Lu, S.M. Pizer, and Sarang Joshi. Principal geodesic analysis for the study of nonlinear statistics of shape. *Medical Imaging, IEEE Transactions on*, 23(8):995–1005, 2004.

- [59] Tom Fletcher, Suresh Venkatasubramanian, and Sarang Joshi. Robust statistics on Riemannian manifolds via the geometric median. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.
- [60] M. Fréchet. Les éléments aléatoires de nature quelconque dans un espace distancié. *Ann. Inst. H. Poincaré*, 10:215–310, 1948.
- [61] M. Fuchs and O. Scherzer. Regularized reconstruction of shapes with statistical a priori knowledge. *International Journal of Computer Vision*, 79(2):119–135, 2008.
- [62] Matthias Fuchs, Bert Jüttler, Otmar Scherzer, and Huaiping Yang. Shape metrics based on elastic deformations. Technical Report 1, 2009.
- [63] Matthias Fuchs and Otmar Scherzer. Segmentation of biologic image data with a-priori knowledge. FSP Report, Forschungsschwerpunkt S92 52, Universität Innsbruck, May 2007.
- [64] H. Garcke. On mathematical models for phase separation in elastically stressed solids, 2000. habilitation thesis.
- [65] Harald Garcke. On Cahn–Hilliard systems with elasticity. *Proceedings of the Royal Society of Edinburgh*, 133(A):307–331, 2003.
- [66] D. Gilbarg and N.S. Trudinger. *Elliptic partial differential equations of second order*. Grundlehren der Mathematischen Wissenschaften. 224. Berlin-Heidelberg-New York: Springer-Verlag, 1992.
- [67] Xu Guo, Kang Zhao, and Michael Yu Wang. Simultaneous shape and topology optimization with implicit topology description functions. *Control and Cybernetics*, 34(1):255–282, 2005.
- [68] B.J. Hafner, S.G. Zachariah, and J.E. Sanders. Characterisation of three-dimensional anatomic shapes using principal components: application to the proximal tibia. *Med. Biol. Eng. Comput.*, 38:9–16, 2000.
- [69] S. Henn and K. Witsch. A multigrid approach for minimizing a nonlinear functional for digital image matching. *Computing*, 64(4):339–348, 2000.
- [70] St. Henn and K. Witsch. Iterative multigrid regularization techniques for image matching. *SIAM J. Sci. Comput.*, 23 No 4:1077–1093, 2001.
- [71] Byung-Woo Hong, Stefano Soatto, and Luminita Vese. Enforcing local context into shape statistics. *Advances in Computational Mathematics*, online first, 2008.
- [72] S. Joshi, B. Davis, M. Jomier, and G. Gerig. Unbiased diffeomorphic atlas construction for computational anatomy. *NeuroImage*, 23:151–160, 2004. Supplement 1.

-
- [73] H. Karcher. Riemannian center of mass and mollifier smoothing. *Communications on Pure and Applied Mathematics*, 30(5):509–541, 1977.
- [74] S. L. Keeling and W. Ring. Medical image registration and interpolation by optical flow with maximal rigidity. *Journal of Mathematical Imaging and Vision*, 23(1):47–65, 2005. to appear.
- [75] D. G. Kendall. Shape manifolds, procrustean metrics, and complex projective spaces. *Bull. London Math. Soc.*, 16:81–121, 1984.
- [76] M. Kilian, N. J. Mitra, and H. Pottmann. Geometric modeling in shape space. In *ACM Transactions on Graphics*, volume 26, pages #64, 1–8, 2007.
- [77] E. Klassen, A. Srivastava, W. Mio, and S. H. Joshi. Analysis of planar shapes using geodesic paths on shape spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(3):372–383, 2004.
- [78] Lew D. Landau and Jewgeni M. Lifschitz. *Hydrodynamik*, volume 6 of *Lehrbuch der theoretischen Physik*. Deutsch (Harri), 1991.
- [79] Michael E. Leventon, W. Eric L. Grimson, and Olivier Faugeras. Statistical shape influence in geodesic active contours. In *5th IEEE EMBS International Summer School on Biomedical Imaging, 2002.*, 2002.
- [80] Z. Liu, J. G. Korvink, and R. Huang. Structure topology optimization: Fully coupled level set method via femlab. *Structural and Multidisciplinary Optimization*, 29:407–417, June 2005.
- [81] J. E. Marsden and T. J. R. Hughes. *Mathematical foundations of Elasticity*. Prentice–Hall, Englewood Cliffs, 1983.
- [82] S. Marsland, C. J. Twining, and C. J. Taylor. Groupwise non–rigid registration using polyharmonic clamped–plate splines. In R. E. Ellis and T. M. Peters, editors, *Medical Image Computing and Computer–Assisted Intervention, MICCAI*, volume 2879 of *LNCS*, pages 771–779, 2003.
- [83] F. Mémoli and G. Sapiro. A theoretical and computational framework for isometry invariant recognition of point cloud data. *Foundations of Computational Mathematics*, 5:313–347, 2005.
- [84] Facundo Mémoli. Gromov-Hausdorff distances in euclidean spaces. In *Workshop on Non-Rigid Shape Analysis and Deformable Image Alignment (CVPR workshop, NORDIA’08)*, 2008.
- [85] Peter W. Michor and David Mumford. Riemannian geometries on spaces of plane curves. *J. Eur. Math. Soc.*, 8:1–48, 2006.

- [86] M. I. Miller, G. E. Christensen, Y. Amit, and U. Grenander. Mathematical textbook of deformable neuroanatomies. *Proc. Natl. Acad. Sci. USA*, 90(24):11944–11948, 1993.
- [87] M. I. Miller and L. Younes. Group actions, homeomorphisms, and matching: a general framework. *International Journal of Computer Vision*, 41(1–2):61–84, 2001.
- [88] M.I. Miller, A. Trouvé, and L. Younes. On the metrics and Euler-Lagrange equations of computational anatomy. *Annual Review of Biomedical Engineering*, 4:375–405, 2002.
- [89] Luciano Modica and Stefano Mortola. Un esempio di Γ^- -convergenza. *Boll. Un. Mat. Ital. B (5)*, 14(1):285–299, 1977.
- [90] J.-M. Morel and S. Solimini. Segmentation of images by variational methods: a constructive approach. *Revista Matemática de la Universidad Complutense de Madrid*, 1(1):169–182, 1988.
- [91] J.M. Morel and S. Solimini. *Variational models in image segmentation*. Birkhäuser, 1995.
- [92] David Mumford and Jayant Shah. Optimal approximation by piecewise smooth functions and associated variational problems. *Communications on Pure Applied Mathematics*, 42:577–685, 1989.
- [93] Pablo Pedregal. *Variational Methods in Nonlinear Elasticity*. SIAM, 2000.
- [94] X. Pennec. Left-invariant Riemannian elasticity: a distance on shape diffeomorphisms? In *Mathematical Foundations of Computational Anatomy - MFCA 2006*, pages 1–14, 2006.
- [95] X. Pennec, R. Stefanescu, V. Arsigny, P. Fillard, and N. Ayache. Riemannian elasticity: A statistical regularization framework for non-linear registration. In *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2005*, LNCS, pages 943–950, 2005.
- [96] Patrick Penzler, Martin Rumpf, and Benedikt Wirth. A phase-field model for compliance shape optimization in nonlinear elasticity. *submitted to ESAIM: COCV*, 2010.
- [97] Dimitrios Perperidis, Raad Mohiaddin, and Daniel Rueckert. Construction of a 4d statistical atlas of the cardiac anatomy and its use in classification. In J. Duncan and G. Gerig, editors, *Medical Image Computing and Computer Assisted Intervention*, volume 3750 of LNCS, pages 402–410, 2005.

- [98] R. D. Rabbitt, J. A. Weiss, G. E. Christensen, and M. I. Miller. Mapping of hyperelastic deformable templates using the finite element method. In *Proc. of SPIE*, volume 2573, pages 252–265, 1995.
- [99] Yogesh Rathi, Samuel Dambreville, and Allen Tannenbaum. Comparative analysis of kernel methods for statistical shape learning. In R.R. Beichel and M. Sonka, editors, *computer vision approaches to medical image analysis*, volume 4241 of *LNCS*, pages 96–107, 2006.
- [100] Yogesh Rathi, Samuel Dambreville, and Allen Tannenbaum. Statistical shape analysis using kernel PCA. In *Proceedings of SPIE*, volume 6064, 2006.
- [101] P. Rogelj and S. Kovačič. Symmetric image registration. *Medical Image Analysis*, 10(3):484–493, 2006.
- [102] D. Rueckert, A. F. Frangi, and J. A. Schnabel. Automatic construction of 3-d statistical deformation models of the brain using nonrigid registration. *IEEE Transactions on Medical Imaging*, 22(8):1014–1025, 2003.
- [103] Martin Rumpf and Benedikt Wirth. An elasticity approach to principal modes of shape variation. In *Proceedings of the Second International Conference on Scale Space Methods and Variational Methods in Computer Vision (SSVM 2009)*, volume 5567 of *Lecture Notes in Computer Science*, pages 709–720, 2009.
- [104] Martin Rumpf and Benedikt Wirth. A nonlinear elastic shape averaging approach. *SIAM Journal on Imaging Sciences*, 2(3):800–833, 2009.
- [105] Martin Rumpf and Benedikt Wirth. An elasticity-based covariance analysis of shapes. *International Journal of Computer Vision*, 2010. accepted.
- [106] Martin Rumpf and Benedikt Wirth. Variational methods in shape analysis. In *Handbook of Mathematical Methods in Imaging*. Springer, accepted, 2010.
- [107] F. R. Schmidt, M. Clausen, and D. Cremers. Shape matching by variational computation of geodesics on a manifold. In *Pattern Recognition*, volume 4174 of *LNCS*, pages 142–151. Springer, 2006.
- [108] Philip Shilane, Patrick Min, Michael Kazhdan, and Thomas Funkhouser. The princeton shape benchmark. In *Proceedings of the Shape Modeling International 2004, Genova*, pages 167–178, 2004.
- [109] O. Sigmund and P. M. Clausen. Topology optimization using a mixed formulation: An alternative way to solve pressure load problems. *Computer Methods in Applied Mechanics and Engineering*, 196(13–16):1874–1889, 2007.

- [110] M. Söhn, M. Birkner, D. Yan, and M. Alber. Modelling individual geometric variation based on dominant eigenmodes of organ deformation: implementation and evaluation. *Phys. Med. Biol.*, 50:5893–5908, 2005.
- [111] C. O. S. Sorzano, P. Thévenaz, and M. Unser. Elastic registration of biological images using vector-spline regularization. *IEEE Transactions on Biomedical Engineering*, 52(4):652–663, 2005.
- [112] Anuj Srivastava, Aastha Jain, Shantanu Joshi, and David Kaziska. Statistical shape models using elastic-string representations. In P.J. Narayanan, editor, *Asian Conference on Computer Vision*, volume 3851 of *LNCS*, pages 612–621, 2006.
- [113] C. Studholme. Simultaneous population based image alignment for template free spatial normalisation of brain anatomy. In J. C. Gee, J. B. A. Maintz, and M. W. Vannier, editors, *Second International Workshop, WBIR, Biomedical Image Registration*, volume 2717 of *LNCS*, pages 81–90, 2003.
- [114] G. Sundaramoorthi, A. Yezzi, and A. Mennucci. Sobolev active contours. *International Journal of Computer Vision.*, 73(3):345–366, 2007.
- [115] V. Šverák. Regularity properties of deformations with finite energy. *Arch. Rat. Mech. Anal.*, 100:105–127, 1988.
- [116] P. M. Thompson and A. W. Toga. *Warping Strategies for Intersubject Registration*, chapter 39, pages 643–674. Academic Press, 2000.
- [117] Nicolas Thorstensen, Florent Segonne, and Renaud Keriven. Pre-image as karcher mean using diffusion maps: Application to shape and image denoising. In *Proceedings of the Second International Conference on Scale Space Methods and Variational Methods in Computer Vision (SSVM 2009)*, volume 5567 of *Lecture Notes in Computer Science*, pages 721–732, 2009.
- [118] A. W. Toga and P. M. Thompson. *Image Registration and the Construction of Multidimensional Brain Atlases*, chapter 43, pages 707–724. Academic Press, 2000.
- [119] Andy Tsai, Anthony Yezzi, William Wells, Clare Tempany, Dewey Tucker, Ayres Fan, W. Eric Grimson, and Alan Willsky. A shape-based approach to the segmentation of medical imagery using level sets. *IEEE Transactions on Medical Imaging*, 22(2):137–154, 2003.
- [120] Gozde Unal, Greg Slabaugh, Anthony Yezzi, and Jason Tyan. Joint segmentation and non-rigid registration without shape priors. Scr-04-tr-7495, Siemens Corporate Research, 2004.
- [121] P. Viola and W.M. Wells. Alignment by maximization of mutual information. *International Journal of Computer Vision*, 24(2):137–154, 1997.

-
- [122] M. Y. Wang, S. Zhou, and H. Ding. Nonlinear diffusions in topology optimization. *Structural and Multidisciplinary Optimization*, 28(4):262–276, 2004.
- [123] Michael Yu Wang, Xiaoming Wang, and Dongming Guo. A level set method for structural topology optimization. *Computer methods in applied mechanics and engineering*, 192:227–246, 2003.
- [124] Michael Yu Wang and Shiwei Zhou. Synthesis of shape and topology of multi-material structures with a phase-field method. *Journal of Computer-Aided Materials Design*, 11(2–3):117–138, 2004.
- [125] Peng Wei and Michael Yu Wang. Piecewise constant level set method for structural topology optimization. *International Journal for Numerical Methods in Engineering*, 78(4):379–402, 2009.
- [126] W. Wells, P. Viola, H. Atsumi, S. Nakajima, and R. Kikinis. Multi-modal volume registration by maximization of mutual information. *Medical Image Analysis*, 1(1):35–51, 1996.
- [127] B. Wirth, L. Bar, M. Rumpf, and G. Sapiro. Geodesics in shape space via variational time discretization. In *Proceedings of the 7th International Conference on Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR’09)*, volume 5681 of *LNCS*, pages 288–302, 2009.
- [128] Benedikt Wirth, Leah Bar, Martin Rumpf, and Guillermo Sapiro. A continuum mechanical approach to geodesics in shape space. *submitted to IJCV*, 2010.
- [129] P. P. Wyatt and J. A. Noble. MAP MRF joint segmentation and registration. In T. Dohi and R. Kikinis, editors, *MICCAI*, volume 2488 of *LNCS*, pages 580–587, 2002.
- [130] Qi Xia and Michael Yu Wang. Simultaneous optimization of the material properties and the topology of functionally graded structures. *Computer-Aided Design*, 40(6):660–675, 2008.
- [131] Qi Xia, Michael Yu Wang, Shengyin Wang, and Shikui Chen. Semi-Lagrange method for level-set based structural topology and shape optimization. *Structural and Multidisciplinary Optimization*, 31(6):419–429, 2005.
- [132] A. Yezzi, L. Zöllei, and T. Kapur. A variational framework for integrating segmentation and registration through active contours. *Medical Image Analysis*, 7(2):171–185, 2003.
- [133] Laurent Younes. Computable elastic distances between shapes. *SIAM J. Appl. Math*, 58(2):565–586, April 1998.

- [134] Yuan-Nan Young and Doron Levy. Registration-based morphing of active contours for segmentation of ct scans. *Mathematical Biosciences and Engineering*, 2(1):79–96, 2005.
- [135] Paul Yushkevich, P. Thomas Fletcher, Sarang Joshi, Andrew Thalla, and Stephen M. Pizer. Continuous medial representations for geometric object modeling in 2d and 3d. *Image and Vision Computing*, 21(1):17–27, 2003.
- [136] Shiwei Zhou and Michael Yu Wang. Multimaterial structural topology optimization with a generalized Cahn–Hilliard model of multiphase transition. *Structural and Multidisciplinary Optimization*, 33:89–111, 2007.
- [137] Lei Zhu, Yan Yang, Steven Haker, and Tannenbaum Allen. An image morphing technique based on optimal mass preserving mapping. *IEEE Transactions on Image Processing*, 16(6):1481–1495, 2007.