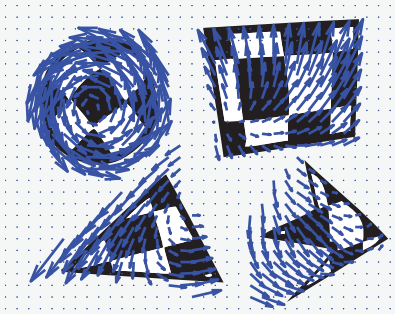
Computer Aided Medical Procedures
Prof. Dr. Nassir Navab



Dissertation

Random Fields for Image Registration

Benjamin M. Glocker

Fakultät für Informatik
Technische Universität München

TECHNISCHE UNIVERSITÄT MÜNCHEN

Computer Aided Medical Procedures & Augmented Reality / I16

Random Fields for Image Registration

Benjamin M. Glocker

Vollständiger Abdruck der von der Fakultät für Informatik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:		Univ.-Prof. Dr. Peter O. A. Struss
Prüfer der Dissertation:		
	1.	Univ.-Prof. Dr. Nassir Navab
	2.	Prof. Dr. Nikos Paragios,
		Ecole Centrale de Paris / Frankreich

Die Dissertation wurde am 09.09.2010 bei der Technischen Universität München eingereicht und durch die Fakultät für Informatik am 16.05.2011 angenommen.

Abstract

Image registration is one of the key components in computer vision and medical image analysis. Motion compensation, multi-modal fusion, atlas matching, image stitching, or optical flow estimation are only some of the applications where efficient registration methods are needed. The task of registration is to recover a spatial transformation which aligns corresponding structures visible in the images. This is commonly formulated as an optimization problem based on an objective function which evaluates the quality of a transformation with respect to the image data and some prior information. So far, mainly classical continuous methods have been considered for the critical part of optimization.

In this thesis, discrete labeling of random fields is introduced as a novel promising and powerful alternative. A general framework is derived which allows to represent both linear and non-linear image registration as labeling problems where random variables play the role of transformation parameters. Based on this framework, several explicit models are defined for the linear and non-linear case. While discrete optimization often provides strong solutions in purely discrete settings, the task of registration actually involves the estimation of continuous transformation parameters. In order to bridge this gap, a novel optimization procedure is proposed based on iterative discrete labeling with successive label space refinement strategies. The procedure is computationally efficient, avoids local minima through large neighborhood search, and yields high-accurate registration.

Besides efficiency, the great advantage of this discrete formulation is that it provides an intuitive control on the search and solution space, prior knowledge can be easily integrated, and it is modular in terms of the objective function since neither numerical nor analytical differentiation is necessary. The implementations are based on the most recent advances in discrete optimization. Performance of the methods is evaluated in numerous medical and non-medical applications such as multi-modal registration, segmentation via atlas matching, deformable image stitching, and optical flow. Experimental results show consistently the great potential of random fields for image registration. This thesis aims at creating a novel and valuable perspective on the modeling part also for other imaging and vision tasks, and hopefully influences the way people think about optimization and the applicability of discrete random fields beyond classical problems.

Keywords:
Image Registration, Markov Random Fields, Discrete Optimization

Zusammenfassung

Die Bildregistrierung ist eine Schlüsselkomponente in Computer Vision und vielen medizinischen Bilderverarbeitungsproblemen. Sei es Bewegungskorrektur, Fusion von multimodalen Bilddaten, nahtloses Aneinanderfügen von Bildern, oder die Berechnung von optischem Fluss, um nur einige Anwendungen zu nennen. Für diese Aufgaben sind effiziente Registrierungsmethoden notwendig. Das Ziel von Registrierung ist die Berechnung einer Bildtransformation, die eine Überlagerung von korrespondierenden Strukturen ermöglicht. Üblicherweise wird Registrierung als Optimierungsproblem formuliert, in welchem eine problemspezifische Zielfunktion die Qualität einer Transformation unter Berücksichtigung der Bilddaten und A-priori-Informationen ermittelt. Zu diesem Zweck werden hauptsächlich klassische kontinuierliche Methoden für den kritischen Teil der Optimierung eingesetzt.

In dieser Dissertation wird diskretes Labeling von Random Fields als neuartige vielversprechende und leistungsfähige Alternative vorgestellt. Ein allgemeines, mathematisches Framework wird erarbeitet, welches erlaubt, sowohl lineare als auch nicht-lineare Registrierung als Labelingproblem zu repräsentieren. Zufallsvariablen übernehmen dabei die Rolle von Transformationsparametern. Basierend auf diesem Framework werden mehrere explizite Modelle für den linearen und nicht-linearen Fall hergeleitet. Während diskrete Optimierung oft zu sehr guten Lösungen in rein diskreten Szenarien führt, beinhaltet die Aufgabe der Bildregistrierung eigentlich die Bestimmung von kontinuierlichen Parametern. Um diese Lücke zu überwinden, wird eine neuartige Prozedur bei der Optimierung vorgeschlagen basierend auf iterativen diskreten Labelings mit schrittweiser Suchraumverfeinerungsstrategie. Die Prozedur ist effizient in der Berechnung, vermeidet lokale Minima durch großräumige Nachbarschaftssuche, und erzielt hochakkurate Registrierung.

Neben der Effizienz besitzt die diskrete Formulierung weitere große Vorteile. Sie bietet eine intuitive Kontrolle über den Such- und Lösungsraum, A-priori-Wissen kann leicht integriert werden, und die Formulierung ist modular im Bezug auf die Zielfunktion, da weder numerische noch analytische Differenzierung benötigt werden. Die in der Arbeit vorgestellten Implementierungen basieren auf den neuesten Entwicklungen in der diskreten Optimierung. Die Performanz der vorgestellten Methoden wird in zahlreichen medizinischen und nicht-medizinischen Anwendungen evaluiert. Darunter sind multi-modale Registrierung, Segmentierung mittels Atlas Matching, deformierbares Stitching, und Berechnung von optischem Fluss. Die Ergebnisse bestätigen ein großes Potential für den Einsatz von Random Fields für die Bildregistrierung. Diese Arbeit soll zusätzlich auch eine neue und wichtige Sichtweise für die Modellierung anderer Bildverarbeitungsprobleme eröffnen und die Anwendbarkeit von diskreten Random Fields über die klassischen Probleme hinaus beeinflussen.

Schlagwörter:
Bildregistrierung, Markov Random Fields, Diskrete Optimierung

CONTENTS

THESIS OUTLINE

In this thesis, we introduce the mathematical framework of random fields and discrete optimization for modeling and solving the problem of image registration. Image registration is one of the key components in computer vision and medical image analysis. The task of registration is to recover a spatial transformation which aligns corresponding structures visible in images. Motion compensation, multi-modal fusion, atlas matching, image stitching, or optical flow estimation are only some applications which rely on efficient registration methods. In many cases, we are interested in extracting high-level information about objects by looking at their motion, deformation, growth or shrinkage over time. In medical applications, the progress of treatments and interventions is often assessed with imaging. In order to see the actual changes of specific regions, for instance a tumor, the images need to be registered. In image-based diagnosis it is often beneficial to fuse images which have been acquired with different sensors or modalities. A non-medical application for image registration is for instance the analysis of video sequences of moving objects observed by a stationary camera. By determining the motion of the individual objects, we can generate further high-level information via motion clustering, classification, or recognition.

Image registration is of interest in many fields and it has been studied by a lot of people in the last decades. Various algorithms have been proposed, some more general and some more application specific methods. In most of these methods, optimization plays an important role. Image registration is commonly formulated as an optimization problem based on an objective function which evaluates the quality of a transformation with respect to the image data and some prior information. So far, mainly classical continuous methods have been considered for the critical part of optimization.

Combinatorial or discrete optimization is a particular sub-field within the huge field of mathematical optimization. Recent advances within this sub-field have lead to an increased popularity of discrete optimization for all kinds of vision and imaging applications. The main contribution of this thesis is to bridge the gap between the discrete world of these powerful optimization methods and the continuous world of image registration. We derive a general framework which allows us to represent both linear and non-linear image registration as discrete optimization problems. We will first give a general introduction into random fields and their particular use in vision and imaging. We will discuss the state-of-the-art optimization methods, but also review some historical methods and the beginnings of the success of random fields. The major part of this thesis is the deriva-

tion and development of models for different tasks of registration based on the random field framework. At the end, we will demonstrate in several applications the promising performance and great potential of our proposed methods.

Here, we give a brief outline and summary of the thesis and the following four chapters.

Chapter 1: Random Fields In our first chapter we will introduce the mathematical, probabilistic framework of random fields which builds the basis for all our approaches and models in image registration. We start by a very basic introduction into graphical models with a focus on popular Markov random fields which have been used immensely in different vision applications. We show the Bayesian justification behind these models, present the famous maximum a posteriori principle, and give a convenient recipe for modeling problems in terms of random fields and energy minimization. Practical considerations are demonstrated with some classical vision applications such as segmentation and stereo matching. We conclude the chapter by discussing some historical background.

Chapter 2: Optimization The second chapter is dedicated to discrete optimization methods. We start by introducing some older methods such as simulated annealing or iterated conditional modes which at their time made it possible at all to consider random fields in practical scenarios. We then focus on state-of-the-art optimization with message passing and graph cuts. Particularly graph cuts are discussed in more detail since our registration methods are all based on optimization via iterative graph-cuts. In this context, we present recent advances in minimizing non-submodular energy functions and first attempts towards hybrid discrete-continuous methods.

Chapter 3: Image Registration Chapter three covers our main contribution, the general framework for registration via discrete labeling of random fields, followed by the derivation of our specific models for the linear and non-linear case. We start with our non-linear models for deformable registration and present an extremely efficient first-order Markov random field approach using free-form deformations. Then we introduce a higher-order Conditional random field based on a piecewise-affine triangulation model. Afterwards, we come to linear registration and present our highly-connected first-order method for this type of registration. We conclude the chapter by a discussion on properties and advantages of discrete formulations, a comparison with related work, and an outlook for possible future work.

Chapter 4: Applications The last chapter of this thesis is dedicated to general experiments and several applications in which we demonstrate the performance and particular properties of our registration methods. We present different medical applications such as rigid multi-modal brain registration, segmentation via atlas matching, and deformable stitching for whole body imaging. Closely related to non-linear registration is the task of optical flow estimation where one seeks to determine the apparent motion of 3D objects in 2D images. We present several experiments and results for our methods when applied to the problem of optical flow.

CHAPTER

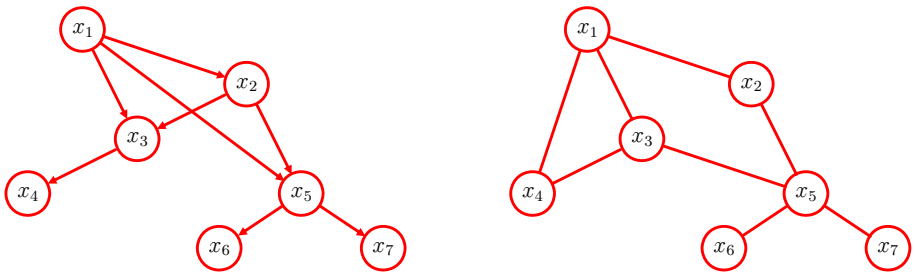
ONE

RANDOM FIELDS

The aim of this chapter is to provide a self-contained introduction to random fields. In particular we focus on the popular Markov random fields and their use in computer vision. We will start by introducing the general concept of probabilistic graphical models of which random fields are a special case. We will then introduce some basics and a consistent notation and terminology which will be used throughout this thesis. The mathematical formulation and probabilistic interpretation of these models is then discussed in detail. Random fields have been successfully applied to many vision problems. We will review some previous works and discuss the motivation and the success of random fields in this area. This introductory chapter should serve as a basis for our approach of image registration using random fields introduced later in Chapter 3. Additionally, we hope that our introduction may be useful to researchers who are interested in applying random fields to other applications.

1.1 Graphical Models

Random fields are a particular class of the so called Probabilistic Graphical Models (PGMs). Before we start to present the concept and mathematical framework behind random fields, it is useful to first have a look at the idea of PGMs. Please note, that parts of this introduction are based on Chapter 8 in [10]. At a first glance, a graphical model is something that we can draw on a piece of paper. It consists of two elementary objects, namely a set of *nodes* V (also called *sites*) and a set of *edges* E (or *links*). These two sets constitute a graph $G = (V, E)$. Nodes represent certain entities of the problem to be modeled and are graphically illustrated as circles. The edges are used to connect nodes and represent some sort of relationship between nodes. Edges can be either *directed* – illustrated as arrows pointing from one node to another – or *undirected* – simply visualized as solid lines. There are two major classes of PGMs, which are the so called Directed Acyclic Graphical Models (DAGMs), and the Undirected Graphical Models (UGMs). Random fields belong to the second class. Two exemplary graphs are shown in Figure 1.1. While DAGMs are suitable for modeling causal relationships (e.g. time-dependent processes) – where the direction of an edge indicates the causality – UGMs



(a) Directed (b) Undirected

Figure 1.1: Variants of graphical models.

are able to model context-dependent relationships between connected nodes. DAGMs do not allow cycles or loops within the graph, since causality with loops is not plausible. In contrast, the graph structure of UGMs is unrestricted allowing to model a broader class of relationships. In practice it depends on the underlying relationship of nodes which of these two models is appropriate for a representation of the problem at hand.

When we talk about node relationship, we should also have a look at what the nodes actual represent. In PGMs every node stands for a *random variable*. That is what brings in the probabilistic meaning into such models. Random variables are the core entities of probability theory which aims at a mathematical understanding and formulation of stochastic processes. An intuitive introduction into probability theory with many examples can be found in [10]. We can think of a random variable as a variable taking certain values from a predefined set of events. The key point is that every event has a certain probability to occur, and with each variable we associate a probability distribution over the set of events. We could for instance model the outcome of a rolling dice as a single random variable where the events are the numbers printed on that dice. Probability theory provides us with a set of rules to compute answers to questions such as: what is the probability of throwing a 6 three times in a row? Indeed, the stochastic process of a rolling dice is rather simple. More complex processes involve a set of random variables and the probabilities of their individual outcome might depend on each other. Random variables might need to "interact" or "communicate" in order to determine their joint probability distribution. PGMs are illustrative and a powerful tool for modeling complex stochastic processes. In addition, PGMs come with a mathematical foundation which allows us to perform probabilistic reasoning on these processes. Following [10], we summarize the main advantages of PGMs, which

1. provide a simple way to *visualize* the structure of stochastic processes, which can be used to *design* and *motivate* new models,

2. allow *insights* by visual *inspection* of the graphs about the relationship (e.g. independence) of random variables,

3. facilitate *complex computations* to be done by simple *graphical manipulations* which perform the operations implicitly.

Probably, the two most popular PGMs are Bayesian networks – which belong to the class of DAGMs – and Markov random fields (MRFs) – belonging to the class of UGMs. In this thesis, we will focus on MRFs and the closely related conditional random fields (CRFs). We will later see why these models are of particular interest in image analysis. More details on Bayesian networks can be found in [10]. We should also note that since all these models are based on the PGM fundament, there exist strong connections between directed and undirected models. In fact, it is possible to transform both into a common form called factor graphs. A unifying view on different models is presented in [139], while some more details on the conversion from one to another can be found in [10].

During the last decade, MRFs have become increasingly popular in all kinds of imaging and vision applications. Their main success is twofold. First, many vision problems can be modeled as an MRF, where the MRF structure is following directly the structure of an image, which makes the model very intuitive. Second, recent advances in MRF optimization algorithms allow efficient computations. A detailed introduction and discussion on MRF optimization is presented in Chapter 2. A motivation for the use of MRFs in computer vision and a brief review of some common applications is later given in Section 1.6. In the following, we will formalize the concept of MRFs and introduce the underlying mathematical framework.

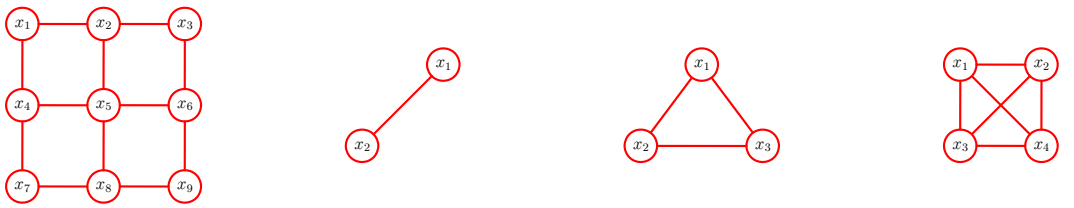
1.2 Markov Random Fields

Let us consider a random field $\mathbf{X}$ which is a set of n random variables $X_i \in \mathbf{X}$. Each variable can take a value $x_i \in L_i$, where L_i is the before mentioned set of events. In random field theory, the events are commonly referred to as labels, and $X_i = x_i$ is then referred to as a label assignment of variable X_i. Note, that in general each variable could have its own predefined set of labels L_i. However, we will see that in many applications, where the variables represent the same type of entity, the variables access a common set L. Once every variable is assigned a label, this is what we call a *labeling* of the field denoted by $\mathbf{X} = \mathbf{x}$ with $\mathbf{x} = (x_1, ..., x_i, ..., x_n)$. Sometimes a labeling is also referred to as a *configuration* or *realization* of the field. The set containing all possible labelings is denoted by $\mathcal{X}$.

As discussed before, we know that an assignment of a label, which equivalently can be seen as the occurrence of a certain event, has a certain probability. We denote this probability by $\rho(X_i = x_i)$ or in brief simply by $\rho(x_i)$. The joint probability of the field labeling is then denoted as $\rho(\mathbf{X}=\mathbf{x})$ or simply as $\rho(\mathbf{x})$.

Now, let us assume the random field is modeling some real world problem in terms of a stochastic process. Naturally, we are interested in computing probabilities such as above, since it would allow us to reason about the process. For instance, we could be interested in a particular probability of a certain event. Or we could ask what is the overall labeling of the field with the highest probability. We will later see examples in vision problems, where exactly these questions arise. There is one special case, where the computation is straightforward. That is when all the random variables are conditionally independent, and we are given the probability distribution associated with each variable.

Simple example: assuming we have four perfect (unbiased) dices. Each dice corre-



(a) Regular grid (b) 1st-order (c) 2nd-order (d) 3rd-order

Figure 1.2: Neighborhoods, cliques, and order.

sponds to a random variable. Our random field is $\mathbf{X} = \{X_1, X_2, X_3, X_4\}$ and the common set of labels is $L = \{1, 2, ..., 6\}$. Without going into more details, we directly know that $\rho(x_i) = 1/6$ for any i and $x_i \in L$, simply because we know that the outcome of a dice is conditionally independent of the other ones, and each value is equally probable. So, we can also easily compute the joint probability $\rho(\mathbf{x}) = \prod_{i=1}^{4} \rho(x_i) = (1/6)^4$.

Obviously, for processes where all the variables are conditionally independent, the concept of random fields is rather useless. But most problems are much more complex and conditional dependence plays a crucial role. We will later see that in many applications the variables – represented by graph nodes – are somewhat located in a spatial domain, for instance equally distributed on a two-dimensional lattice. Here, conditional dependence between neighboring nodes is of particular interest. Modeling these dependencies is one of the strength of random fields. We already learned that node relationships can be represented via graph edges. Let us now discuss this concept of relationship in more detail, before we present the properties of a special class of random fields, the Markov random fields.

1.2.1 Neighborhoods, Cliques, and Order

The graph nodes V have a one-to-one correspondence to the set of random variables $\mathbf{X}$, i.e. each node $i \in V$ is associated with a variable $X_i \in \mathbf{X}$. The edges in the corresponding graph of a random field define a neighborhood system on the set of nodes. To illustrate this, let us assume the nodes are distributed on a regular two-dimensional lattice. Then a simple 4-connected neighborhood system corresponds to what we call a regular grid where inner nodes have exactly four neighbors which are the adjacent nodes on the left, right, top, and bottom. Border nodes will have three neighbors, while the four corner nodes will have only two. Such a graph is illustrated in Figure 1.2(a). This leads us to the definition of a *clique*. A clique is a subset of nodes $C \subseteq V$. If $|C| > 1$, every node $i \in C$ has to be a direct neighbor to all other nodes $j \in C$ (i.e. $(i, j) \in E$). So a clique is either a single node, or it constitutes a fully-connected subgraph. For the regular grid we see that it contains cliques of size up to two.

In general, the total number of cliques of a random field is equal to the number of fully-connected subgraphs plus the number of nodes. We define the set $\mathbf{C}$ to be the set containing all cliques of a random field. We further define the *order* of a random field as the size of the maximal clique minus one. This means, a first-order random field contains

only cliques of size up to two (cf. Figure 1.2(b)), a second-order field contains cliques of size up to three (cf. Figure 1.2(c)), and so on. The fully-connected graph shown in 1.2(d) corresponds to a third-order clique. In general, for fully-connected graphs the set of cliques $\mathbf{C}$ is equal to the power set of V.

Sometimes, we need to refer to all neighbors of one specific node. To this end, we define a neighborhood set N_i which contains all nodes connected to i. By definition node i is not a neighbor of itself. A neighborhood system of a random field is then defined as $N = \{N_i \,|\, \forall i \in V\}$. As mentioned above, the neighborhoods reveal something about the conditional dependency between random variables. This will be discussed in the following.

1.2.2 Markov Properties and Local Characteristics

Based on the above definition of neighborhoods, we can now introduce the properties of Markov random fields.

Definition 1 (Markov random field) *A random field* $\mathbf{X}$ *is said to be a Markov random field with respect to a neighborhood system* N *if and only if it satisfies the following two properties:*

$$\rho(\mathbf{x}) > 0 \quad \forall\, \mathbf{x} \in \mathcal{X} \quad , \qquad \textit{(Positivity)} \tag{1.1}$$

$$\rho(x_i | \{x_j : j \in N_i\}) = \rho(x_i | \{x_j : j \in V \setminus \{i\}\}) \quad . \qquad \textit{(Markovian)} \tag{1.2}$$

The first property ensures that the joint probability can be uniquely determined by requiring that any labeling has a strictly positive probability. This property is usually satisfied in practice, or can be easily ensured. The second property states something about the conditional dependence between neighboring nodes, which yields two interesting observations about MRFs. First, Equation (1.2) tells us that any node depends only on its direct neighbors. This depicts the local characteristics of MRFs. Second, if two nodes are not connected it automatically implies that these nodes are conditionally independent. A detailed discussion of these two properties can found in [5].

While the theoretical implication of locality of Equation (1.2) is of great importance, specifying the MRF in terms of conditional probabilities can be quite difficult. The problem is that conditional probabilities are subject to some non obvious and highly restrictive consistency conditions [5, 102]. The fact that the nodes in a random field are in general unordered and no hierarchy exists adds to this difficulty. It is not obvious, how to deduce the joint probability from the conditional probabilities. This is in contrast to Bayesian networks, where the factorization of the joint probability in terms of conditional probabilities is straightforward given the natural hierarchy of nodes represented by directed edges. However, there is another, more intuitive way to specify an MRF directly in terms of the joint probability. In the following, we show how this can be done.

1.2.3 Markov-Gibbs-Equivalence

The specification of an MRF via its joint probability is based on a theoretical result known as the Hammersley-Clifford-Theorem [51]. But first, we need to define another class of random fields, the so called Gibbs random fields (GRFs).

Definition 2 (Gibbs random field) *A random field* $\mathbf{X}$ *is said to be a Gibbs random field if and only if its joint distribution* $\rho(\mathbf{x})$ *is a Gibbs distribution, which has the following form:*

$$\rho(\mathbf{x}) = \frac{1}{Z} \prod_{C \in \mathbf{C}} \exp\left(-\psi_C(\mathbf{x})\right) \quad \text{with} \quad Z = \sum_{\mathbf{x} \in \mathcal{X}} \prod_{C \in \mathbf{C}} \exp\left(-\psi_C(\mathbf{x})\right) \quad , \tag{1.3}$$

where Z is a normalization constant (also known as the partition function) which ensures that the factors sum up to one and constitute a probability distribution. The functions $\psi_C(\mathbf{x})$ are the so called *potential functions*, where one function is defined per clique $C \in \mathbf{C}$. So, for an GRF the joint probability is simply factorized into terms based on the exponential of the negative values of the potential functions. Later, we will see what these functions represent. The definition of these functions is an essential part of modeling any problem as a random field. For now, it is sufficient to keep in mind that these are unrestricted real-valued functions which evaluate sub-labelings in terms of associated energies, where the lower the energy the more likely the labeling. Gibbs distributions arise for instance in statistical mechanics as the equilibrium distribution of a system with energy function ψ [37].

The following important theorem states the equivalence between MRFs and GRFs. Proofs can be found in [51] and [5].

Theorem 1 (Hammersley-Clifford) *A random field* $\mathbf{X}$ *is said to be a Markov random field if and only if* $\mathbf{X}$ *is a Gibbs random field, and vice versa.*

Having this in mind, we are able to specify the joint probability of an MRF by specifying the potential functions of a Gibbs distribution.

1.3 Modeling with MRFs

In the following, we will consider an intuitive example – the task of image restoration – to illustrate the whole process of MRF modeling. In image restoration we are given a noisy version of an image and the goal is to restore the original intensities of each pixel. This example is inspired by the seminal work on MRFs in vision by Geman and Geman [36].

The first step in MRF modeling is to define the role of the random variables. Let us introduce two sets of variables $\mathbf{X}$ and $\mathbf{Y}$. Assuming the discrete image I contains n pixels, then the two sets of variables have both the cardinality $n = |\mathbf{X}| = |\mathbf{Y}|$. The set $\mathbf{X}$ corresponds to the pixels of the restored image and their values correspond to the restored intensities. The set $\mathbf{Y}$ is associated with the noisy image and their values correspond to the noisy intensities. The values of $\mathbf{Y}$ are given or *observed*; we call these fixed values the *observation* (cf. Figure 1.3(b)). In contrast, the values of $\mathbf{X}$ are unknown or *hidden*. These are the values we want to estimate. Estimating the hidden variables is also called an *inference* problem.

Now, we come to the essential point in modeling. We need to make suitable assumptions, simplifications, and/or approximations about the problem at hand. Here, we make the following two assumptions. First, we say that we believe that the original intensities

are related to the observed ones, i.e. the restored intensities should be somehow similar to the ones in the noisy image. If this would not be valid, there is little chance of restoring the image. The second assumption is that neighboring pixels most probably have similar intensities. This seems valid for a large portion of the image, except for the boundary pixels between different objects.

From these assumptions we can deduce the relationships between variables. We formalize this by defining a Markov random field on the set $\mathbf{X} \bigcup \mathbf{Y}$. We introduce edges between pairs (X_i, Y_i) which define the conditional dependence between hidden and observed variables. Assuming that the variables $\mathbf{X}$ are spatially distributed on a two-dimensional lattice (following the natural grid structure of the discrete image I), we further define a 4-connected neighborhood system on $\mathbf{X}$. Every hidden variable is connected to its adjacent hidden variables (cf. Figure 1.2(a)). This formalizes our second assumption about the conditional dependence between neighboring pixels. The corresponding first-order MRF is illustrated in Figure 1.3(a). A configuration of the MRF is represented by the labeling $(\mathbf{x}, \mathbf{y})$, where $\mathbf{y}$ is fixed to the observed values and $\mathbf{x}$ corresponds to a restored image. The joint distribution of the above MRF can be written in form of a Gibbs distribution:

$$\rho(\mathbf{x}, \mathbf{y}) = \frac{1}{Z} \prod_{C \in \mathbf{C}} \exp\left(-\psi_C(\mathbf{x}, \mathbf{y})\right) \quad . \tag{1.4}$$

Given any image, the question of interest would be: what is the probability of that image to be a good restoration given the observation? Or mathematically, we seek the conditional probability distribution $\rho(\mathbf{x}|\mathbf{y})$. Computing this distribution is not easy, often infeasible. Why? Because there is most probably an infinite number of observations, i.e. noisy images; all of them have to be known in order to define the distribution properly. In the following, we will see that it is not necessary to know the exact conditional distribution. We are still able to reason about the hidden variables based on the joint distribution in Equation (1.4).

1.3.1 Posterior, Likelihood, and Prior

The rules of probability allow us to derive a connection between the conditional probability and the joint probability, i.e. $\rho(\mathbf{x}, \mathbf{y}) = \rho(\mathbf{x}|\mathbf{y})\rho(\mathbf{y})$. Bringing $\rho(\mathbf{y})$ to the other side, we get $\rho(\mathbf{x}|\mathbf{y}) = \rho(\mathbf{x}, \mathbf{y})/\rho(\mathbf{y})$. Using the symmetry property of the joint distribution, we can derive a very well known rule of probability, namely the Bayes' Theorem:

$$\rho(\mathbf{x}|\mathbf{y}) = \frac{\rho(\mathbf{y}|\mathbf{x})\rho(\mathbf{x})}{\rho(\mathbf{y})} \quad . \tag{1.5}$$

For the term $\rho(\mathbf{y})$ we can safely assume a constant factor. It seems reasonable to assume that any observation has same probability. We get $\rho(\mathbf{x}|\mathbf{y}) \propto \rho(\mathbf{x}, \mathbf{y}) = \rho(\mathbf{y}|\mathbf{x})\rho(\mathbf{x})$, which states that the Gibbs distribution in Equation (1.4) is proportional to the conditional distribution. This is of great importance in the following section. Bayes' terminology introduces special names for the single probability distributions, i.e.

- *Prior distribution:* $\rho(\mathbf{x})$

- *Likelihood distribution:* $\rho(\mathbf{y}|\mathbf{x})$
- *Posterior distribution:* $\rho(\mathbf{x}|\mathbf{y})$

The prior distribution $\rho(\mathbf{x})$ reflects *a priori* knowledge about the hidden variables independently of the observation $\mathbf{y}$. This knowledge is available before we obtain any observation. For instance, the assumption that neighboring pixels should have similar intensities is a so called *prior*. Such priors impose constraints on the solution space of $\mathbf{x}$. If no prior information is available, one assumes an uniform distribution, where every labeling has equal prior probability. The likelihood distribution $\rho(\mathbf{y}|\mathbf{x})$ evaluates how well a certain labeling of the hidden variables fits the observation. In our example, the assumption that the restored intensities should be somehow similar to the observed ones is such a *likelihood*. The posterior distribution $\rho(\mathbf{x}|\mathbf{y})$ reflects the probability of a labeling after we have made an observation and when we combine the prior and the likelihood. The connection between these three distributions was already shown above, i.e. $\textit{posterior} \propto \textit{likelihood} \times \textit{prior}$. If we find a labeling that maximizes the right hand side, this labeling will also have the maximum posterior probability.

Before we discuss this in more detail, let us decompose Equation (1.4) into the likelihood and prior distribution. For the restoration problem, we have defined two types of cliques in $\mathbf{C}$ where all cliques have size two. Let us denote the potential functions for the (X_i, Y_i) connections simply by $\psi_i(\mathbf{x}, \mathbf{y})$. Further, we denote the potential functions for the connections between adjacent hidden variables (X_i, X_j) by $\psi_{ij}(\mathbf{x}, \mathbf{y})$. Additional simplifications can be done on the argument of these functions. ψ_i considers only the two labels x_i and y_i, while ψ_{ij} considers only x_i, x_j and is completely independent of the labeling $\mathbf{y}$. A graphical representation of these terms in form of edges in a random field is shown in Figure 1.3(a). We can rewrite the Gibbs distribution as a product of two separate parts

$$\rho(\mathbf{x}, \mathbf{y}) = \frac{1}{Z} \prod_i \exp\left(-\psi_i(x_i, y_i)\right) \prod_{(i,j)} \exp\left(-\psi_{ij}(x_i, x_j)\right) \quad . \tag{1.6}$$

The first part of the product corresponds to the likelihood distribution

$$\rho(\mathbf{y}|\mathbf{x}) = \frac{1}{Z_i} \prod_i \exp\left(-\psi_i(x_i, y_i)\right) \quad , \tag{1.7}$$

while the second part, which is independent of the observation, corresponds to the prior

$$\rho(\mathbf{x}) = \frac{1}{Z_{ij}} \prod_{(i,j)} \exp\left(-\psi_{ij}(x_i, x_j)\right) \quad . \tag{1.8}$$

As we will see, this is a quite convenient view on the joint distribution. It allows us to separately define suitable potential functions for the problem at hand, which then fully define the joint distribution. Let us now try to encode our rather vague assumptions for the restoration process in clearly defined real-valued potential functions. For the likelihood term, we assume that the noise in the observed image follows a normal distribution $\mathcal{N}(0, 1)$

with zero mean and a standard deviation of one. Therefore, a suitable potential function for the likelihood is

$$\psi_i(x_i, y_i) = (x_i - y_i)^2 \quad , \tag{1.9}$$

which is simply the squared difference of intensities. For the prior, we use a function which preserves discontinuities. The corresponding potentials are defined as

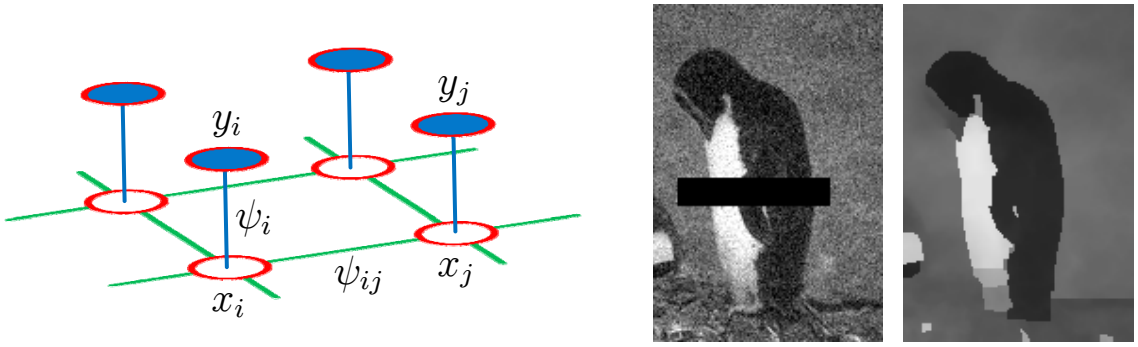
$$\psi_{ij}(x_i, x_j) = \min\left[(x_i - x_j)^2, t\right] \quad , \tag{1.10}$$

where t is a threshold on the maximum penalty, which might have to be adjusted to the given image data. The motivation for this truncated function comes from the fact, that we assume similar values for neighboring pixels, but we also know that discontinuities in image intensities occur at object boundaries. Such discontinuities should not be over-penalized. The interested reader can find an intuitive introduction and formal mathematical derivation of such discontinuity preserving priors in [11]. We borrowed the definitions of the potential functions from [143]. Note that, for simplicity, very often the explicit definition of the observed variables $\mathbf{y}$ is omitted and thus these variables will also not occur in the argument of the potential functions. For instance, the likelihood potentials are then simply written as $\psi_i(x_i)$, and the dependence on the observation is assumed implicitly. Also worth to note, the potential functions have special names related to on how many hidden variables they depend. The ψ_i functions are then called *unary potentials*, while we denote the ψ_{ij} terms as *pairwise potentials*.

As mentioned earlier, we are interested in labeling the hidden variables such that the posterior probability is maximized. Assuming that we can find this labeling, our hope is that this labeling corresponds to the desired solution, i.e. in case of restoration, this labeling should correspond to a visually good and reasonable looking image. In Chapter 2, we will present algorithms which allow us to compute these labelings. In the following, we will introduce the concept behind this idea from a probabilistic perspective. But let us first briefly introduce another class of random fields closely related to MRFs.

1.3.2 Conditional Random Fields

Now, that we have seen an example for the posterior distribution of an MRF, we can discuss another class of random fields, the so called conditional random fields. The main difference between MRFs and CRFs is how the observation $\mathbf{y}$ is integrated in the potential functions. In the restoration example, we have seen that each unary potential ψ_i is a function of the hidden variable x_i and the observation y_i, while each pairwise potential ψ_{ij} is independent of the observation. This is different in CRFs where all potentials are functions of the whole observation [102]. We also say that a CRF is globally conditioned on the observation. CRFs for instance allow to encode data-dependent priors or higher-order likelihood terms, i.e. likelihood terms which consider a larger set of observed variables. In the literature, at least in the vision and imaging communities, the differentiation between MRFs and CRFs is often not very clear; both terms are used interchangeably. It is good to know that there is a difference and usually we can easily see whether a model belongs to the class of MRFs or CRFs by careful inspection of the potential functions. Whenever



(a) 1st-order MRF model (b) Observation (c) MAP estimate

Figure 1.3: Image Restoration. The random field is shown in (a). Blue edges represent the likelihood terms depending on the hidden state x_i and the observation y_i. Green edges represent the prior terms depending on two neighboring hidden states x_i and x_j. An exemplary observed image (taken from [143]) is shown in (b) which corresponds to the fixed labeling $\mathbf{y}$. The MAP estimate $\hat{\mathbf{x}}$ is shown in (c). The black area in (b) is corrupted and no observation is available. In this area, all likelihood terms are set to zero and the resulting labels in $\hat{\mathbf{x}}$ come from the prior.

we derive a random field model in the following, we will try to emphasize whether it is an MRF or CRF. However, in the more general paragraphs where we explain some of the basic concepts we might tend to use only the term Markov random field even in the case where the claims or assumptions are valid for both classes.

1.4 Maximum A Posteriori

In the preceding sections, we modeled the problem of image restoration by means of a Markov random field. We first defined the role of the random variables and their interrelationships. By defining the associated potential functions, the MRF is fully specified. Now, in order to compute a solution to the problem, i.e. a restored version of the observed image, we seek to maximize the posterior probability. Mathematically, this can be written as

$$\hat{\mathbf{x}} = \arg\max_{\mathbf{x}} \rho(\mathbf{x}|\mathbf{y}) \quad . \tag{1.11}$$

Here, $\hat{\mathbf{x}}$ is the labeling which maximizes the posterior probability $\rho(\mathbf{x}|\mathbf{y})$. This labeling is called the *maximum a posteriori* (MAP) estimate. We have already seen that $\rho(\mathbf{x}|\mathbf{y}) \propto \rho(\mathbf{y}|\mathbf{x})\rho(\mathbf{x})$. Thus, the MAP estimate can be equivalently found by

$$\hat{\mathbf{x}} = \arg\max_{\mathbf{x}} \rho(\mathbf{y}|\mathbf{x})\rho(\mathbf{x}) \quad . \tag{1.12}$$

To get a little bit more practical, let us insert the likelihood and prior distributions of our restoration example:

$$\hat{\mathbf{x}} = \arg\max_{\mathbf{x}} \frac{1}{Z} \prod_i \exp\left(-\psi_i(x_i, y_i)\right) \prod_{(i,j)} \exp\left(-\psi_{ij}(x_i, x_j)\right) \quad . \tag{1.13}$$

In order to calculate the probability distributions, we also have to evaluate Z. However, the computation is usually intractable considering its definition in Equation (1.3). Fortunately, when we are only interested in the MAP estimate, the explicit evaluation of Z can be avoided by a simple mathematical trick. We can convert the maximization into an equivalent minimization problem and get rid of the probability distributions. This is demonstrated in the following section.

On a side note, if the prior distribution is flat, i.e. a uniform distribution which is assumed when no prior information is available, the MAP estimate is equivalent to the *maximum likelihood estimate* (MLE).

1.4.1 Energy Formulation

Given a Gibbs distribution, we can extract the corresponding Gibbs energy by taking the negative logarithm:

$$\begin{aligned}\mathcal{E}(\mathbf{x}) &= -\log\left(\rho(\mathbf{x})\right) \\ &= -\log\left(\frac{1}{Z}\prod_{C\in\mathbf{C}}\exp\left(-\psi_C(\mathbf{x})\right)\right) \\ &= -\log\left(\frac{1}{Z}\right) + \sum_{C\in\mathbf{C}}\psi_C(\mathbf{x}) \\ &= \text{const} + \sum_{C\in\mathbf{C}}\psi_C(\mathbf{x}) \quad .\end{aligned} \tag{1.14}$$

We see that the energy is the sum over the potential functions plus a constant value. This observation allows us to reformulate the MAP estimation in Equation (1.11) in terms of an energy minimization problem:

$$\begin{aligned}\hat{\mathbf{x}} &= \arg\min_{\mathbf{x}}\mathcal{E}(\mathbf{x}|\mathbf{y}) \\ &= \arg\min_{\mathbf{x}}\mathcal{E}(\mathbf{y}|\mathbf{x}) + \mathcal{E}(\mathbf{x}) \quad ,\end{aligned} \tag{1.15}$$

where $\mathcal{E}(\mathbf{x}|\mathbf{y})$ is called the *posterior energy*, $\mathcal{E}(\mathbf{y}|\mathbf{x})$ is the *likelihood energy*, and $\mathcal{E}(\mathbf{x})$ is the *prior energy*. If we insert the likelihood and prior of our restoration example, we get

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}}\sum_i \psi_i(x_i, y_i) + \sum_{(i,j)}\psi_{ij}(x_i, x_j) + \text{const} \quad . \tag{1.16}$$

The constant value can be safely omitted, since it does not change the localization of the minimum. Energy minimization is a very common and convenient approach for solving problems such as the restoration of images. In general, we believe it is much more intuitive to think in terms of energies than in terms of probabilities. Remember, when we defined the potential functions for the image restoration; the idea that the hidden variables should be somewhat close to the observation can be directly expressed by the squared differences as in Equation (1.9). There is no need to think too much about the underlying probabilistic meaning.

The field of optimization provides many algorithms which can be directly used to compute the MAP estimate by energy minimization. We will discuss some of these algorithms in detail later in Chapter 2. Just to provide an intuition at this point how we could perform the optimization, consider a simple *direct descent* approach (see e.g. Chapter 3 in [11]). Direct descent runs like this: iterate over the set of random variables; in each iteration propose a local change for the random variables X_i; check whether the energy decreases; if it does then make the change. After several cycles, this algorithm will converge to an energy minimum. However, this will be in general a local one and not necessarily a good one. We will later see algorithms which can perform much better.

1.5 MRF Recipe

It is time to sum up a little bit what we have seen so far. We have learned how we can pose a vision problem in terms of a random field. We first identified the entities of the problem and introduced the corresponding random variables. We then defined a set of labels – the values which can be assigned to the variables. We have chosen a graph topology specifying the conditional dependency between variables. Suitable assumptions about the problem at hand were made through the definition of a likelihood and prior both encoded in the potential functions of a Gibbs distribution. We have further seen, that a solution to our problem can be computed via MAP estimation by solving an energy minimization problem. We can summarize these steps in a general MRF recipe [102]:

1. Pose a vision problem as a labeling problem.
   (a) Identify the role of random variables and labels.
   (b) Set up the MRF graph with a neighborhood system (edges and cliques).
2. Derive the posterior energy $\mathcal{E}(\mathbf{x}|\mathbf{y})$ that fits to the problem at hand.
   (a) Derive the likelihood energy $\mathcal{E}(\mathbf{y}|\mathbf{x})$.
   (b) Derive the prior energy $\mathcal{E}(\mathbf{x})$.
3. Find the MAP solution via energy minimization.

The last point will be discussed in Chapter 2. In the following section we would like to present some classical vision applications which have been posed as discrete labeling problems – where the random variables take values from a discrete set. The great success of discrete labeling in these applications is mainly due to efficient algorithms for computing the MAP solution.

1.6 Discrete Labeling in Computer Vision

Labeling problems are categorized into discrete and continuous ones, depending on whether the set of labels is finite and discrete or infinite and continuous. In the discrete case, the set of labels can be represented by a set of integer values as $L = \{1, ..., N\}$,

while in the continuous case the label set is $L \subseteq \mathbb{R}$. Note, that in case of discrete sets, the labels can have a symbolical meaning or the integer values map to a quantized set of real values. At a first glance, the question whether a problem should be modeled as a discrete or continuous random field seems rather easy to answer. Of course, it highly depends on the nature of the entities represented by the random variables. Considering the previous example of modeling the outcome of a dice, it is clear that this is a naturally discrete setting. If a random variable is modeling the temperature distribution in some physical process, a continuous representation seems to be the right choice. So what about the restoration problem where the random variables stand for the restored intensities in an image? Ideally, the range of intensities is continuous, but since we are processing the images digitally, we also deal with quantized ranges of intensity. For instance, a simple gray-scale image with an 8-bit encoding allows intensity values within a discrete range of $[0, 255]$. A discrete random field seems sufficient.

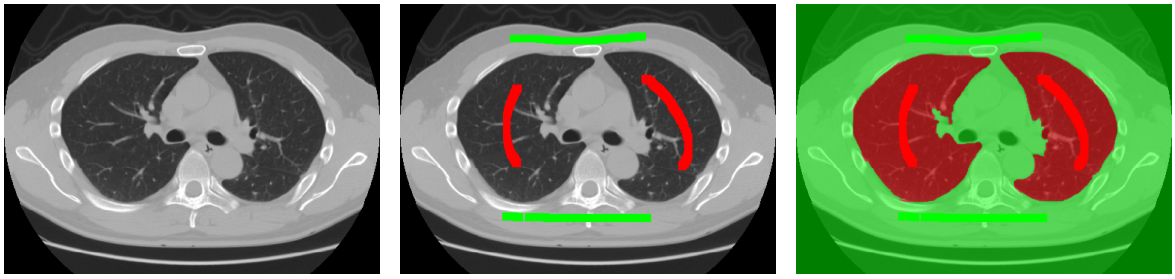
In computer vision, we find many problems which have been addressed by discrete random fields. Some of them are of discrete nature while others have been explicitly transformed into discrete problems by employing suitable approximations and/or simplifications. The main motivation for modeling a problem right away in a discrete setting is that it allows to make use of recent powerful discrete or combinatorial optimization. One might argue that for some cases a continuous representation is actually the better choice – which might be true – but even then, one has to keep in mind that for any problem which is solved on a computer discretization is unavoidable at some point. The only question is: when and where do we have to do this discretization?

One claim of this thesis will be that even for highly continuous problems such as image registration where random variables will represent transformation parameters, an early discretization – already in the phase of modeling – can yield extremely efficient and accurate registration algorithms. This will be later discussed in Chapter 3. Let us now have a look at two classical vision problems – beyond the restoration problem – successfully solved through discrete random fields. These two examples, and some more, can be also found in [143].

1.6.1 Segmentation

Image segmentation is the task of extracting an object from a scene. From a low-level perspective, this can be achieved by labeling individual image points to belong either to the object or to the background, where the latter one compasses all pixels not belonging to the object to be segmented. This task can be naturally formulated as a binary labeling problem, where the set of labels is simply $L = \{0, 1\}$ (or symbolically $L = \{\text{“bkg”}, \text{“obj”}\}$) and every pixel is represented by one random variable; a common 4-connected neighborhood can be considered. A seminal work on binary segmentation formulated as a first-order MRF and solved by discrete optimization is [15]. The likelihood energy is encoded in the unary potentials as

$$\psi_i(x_i) = \begin{cases} -\log \rho(d_i \,|\, \text{“bkg”}) & \text{if } x_i = 0 \\ -\log \rho(d_i \,|\, \text{“obj”}) & \text{otherwise} \end{cases} . \tag{1.17}$$



(a) CT scan (b) User marks (c) Segmentation

Figure 1.4: Image segmentation. The image to be segmented is shown in (a). In (b) a user has marked representative regions for the object (lung tissue) in red, and the background (all other tissue) in green. In (c) the resulting binary labeling corresponding to the lung segmentation.

The likelihood energy makes use of predetermined probability distributions. If we want to assign a certain label x_i, for instance the background label "bkg", to an image pixel i, the unary terms evaluate how likely that label is with respect to the pixel's intensity value d_i. The idea is, that the user interactively marks small representative regions for the background and the object beforehand from which the distributions are determined (cf. Figure 1.4(b)). Additionally, a prior similar to the restoration example is encoded on the pairwise potentials as

$$\psi_{ij}(x_i, x_j) = \exp\left(-\frac{(d_i - d_j)^2}{2\sigma^2}\right) \cdot \frac{1}{\|i - j\|} \cdot \delta(x_i, x_j) \quad , \tag{1.18}$$

with

$$\delta(x_i, x_j) = \begin{cases} 0 & \text{if } x_i = x_j \\ 1 & \text{otherwise} \end{cases} \quad . \tag{1.19}$$

The prior favors equal labels for neighboring pixels by assigning zero cost if two labels x_i and x_j are the same. The cost for assigning different labels – corresponding to a segmentation boundary between pixels i and j – depends on the intensity difference of these two pixels. The (contrast-sensitive) cost is decreasing with an increasing difference of intensities. This encodes the idea, that the pixels of the object to be segmented have similar appearance. The prior cost also depends on the Euclidean distance of pixels, which in the case of a 4-connected neighborhood is of course constant. The parameter σ acts as a noise parameter. A small σ allows less variation of the intensity values within the object, while a larger σ should be chosen when the image exhibits a considerably amount of noise. The above definitions only sketch the basic principle of MRF-based binary segmentation, also called graph-cut segmentation. An example is illustrated in Figure 1.4. A more recent extension of this approach is presented in [127].

1.6.2 Stereo Matching

Another classical example for a vision problem casted as a discrete MRF is stereo matching [17, 145]. Here, the task is to determine point correspondences between two images both

(a) (b) (c)

Figure 1.5: Stereo matching. The two input images in (a) and (b). The dense disparity map in (c).

showing the same scene at the same time but from different views. It is assumed that the relationship between the two views is known; the stereo camera system is calibrated. The human vision system is a biological example for such a calibrated system. Stereo systems allow to infer three-dimensional information about a scene. The depth or distance of an object to the observer is proportional to its disparity observable in the two views. Through identification of point correspondences in the images, we can determine these disparities and compute a dense depth map via triangulation. The nice thing in stereo matching is that the images can be preprocessed (i.e. rectified) such that the search for correspondences is restricted to one dimension (along the epipolar lines [53]). An example for stereo matching is shown in Figure 1.5.

In order to formulate the stereo matching as a discrete labeling problem, we assume a finite number of depth layers. So the set of labels can be defined as $L = \{0, ..., N\}$, i.e. a discrete set of potential disparities. Again, every pixel is a random variables and a 4-connected neighborhood can be considered. Let $I_i(p)$ be the function determining the intensity value of image i at point p, we can define the likelihood energy as

$$\psi_i(x_i) = |I_1(p_i + x_i) - I_2(p_i)| \quad . \tag{1.20}$$

The unary potentials are measuring the photo-consistency of an image point p_i in one image with the displaced point $p_i + x_i$ in the other by means of simple intensity difference. Again, the same pairwise prior as in Equation 1.10 can be used to favor similar disparities between neighboring pixels, while preserving discontinuities at depth boundaries.

The above definitions present only the basic principle behind solving the stereo problem by means of discrete random fields. More sophisticated energy terms – in particular for the likelihood – are used in practice which are more robust to noise and outliers. An overview of state-of-the-art algorithms and further examples such as shown in Figure 1.5 can be found on the website[1] of the Middlebury stereo database described in [134].

[1] `http://vision.middlebury.edu/stereo/`

1.7 Historical Notes and Related Perspectives

It is quite interesting to have a look at the history of random fields and their first application to imaging problems. Spatial, contextual interactions on lattice-like graphs have a broad range of applications in various fields of statistical science. The origin of today's MRF framework can be dated back to physics. In the early 1920s, Ernst Ising, a German physicist and student of Wilhelm Lenz, developed a mathematical model for ferromagnetism in solid state bodies. Ising defined a set of nodes equally distributed on a rectangular domain; each node corresponds to a dipole which at any given moment is in one of two states, "up" or "down". He derived the probabilities for the configurations of the field to be given by a Gibbs distribution. Today, we know that this is equivalent to an MRF binary labeling problem. Kindermann and Snell [69] provide an excellent introduction to MRFs and devote a whole chapter on the history of the Ising model.

Definitely, the proof of equivalence of Gibbs random fields and Markov random fields – first presented in the unpublished work of Hammersley and Clifford [51] – contributed a lot to the popularity of random field theory. Thanks to this, we have a profound framework which allows to define, determine, manipulate, and infer the underlying probability distributions in a convenient way. Since the early 1970s, discrete MRFs found their way into the field of engineering as an important tool for modeling, introduced by John Woods [159]. In the foreword of Stan Li's recent book on MRFs in image analysis [102], Rama Chellappa states: "A big impetus to theoretical and practical considerations of 2D spatial interaction models, of which MRFs form a subclass, was given by the seminal works of Julian Besag [5, 6]. Since the early 1980s, MRFs have dominated the fields of image processing, image analysis and computer vision".

Without doubt, one of the most inspiring papers on MRFs in vision is the work of Geman and Geman [36]. They were the first who tackled a vision problem – the restoration of noisy images – by means of an MRF formulation. Their work had a huge impact on the following twenty-five years in this field. Two comprehensive sources on MRFs and its early application to image processing and computer vision are the overview article by Geman and Graffigne [37] and the book edited by Chellappa and Jain [22]. Another excellent book by Blake and Zissermann [11] provides an intuitive perspective on the principles of MRFs. The authors introduce random fields as so called cooperative networks where cells communicate and exchange information. Their illustrative examples – such as the wire frame covered by a soap film, the system of springs representing the energy function, and in particular the outbreak of the Cabbage Mosaic Virus on a field of cabbages arranged in a regular tessellation – make the MRF concept very clear, intuitive, and quite amusing.

Modeling vision and imaging problems in terms of random fields is now around for while, almost 30 years. Gradually, some fundamental, intuitive and some quite complex models have been derived for all kind of theoretical problems. But often the complexity and computational costs hindered the move towards practical solutions. A fortiori, it is very exciting to observe an immense push within this field since the end of the 1990's when more and more efficient discrete optimization techniques became available. Thanks to graph-cuts [17] and efficient message passing algorithms [35], random fields have become ubiquitous and conquered in particular the field of image segmentation. Nowadays, the

computational burden which usually comes along with higher-order models has reduced significantly. It is time to move one step further and apply this powerful framework to problems which, at all appearances, are not necessarily suitable to be fitted into this field. Problems such as image registration.

CHAPTER

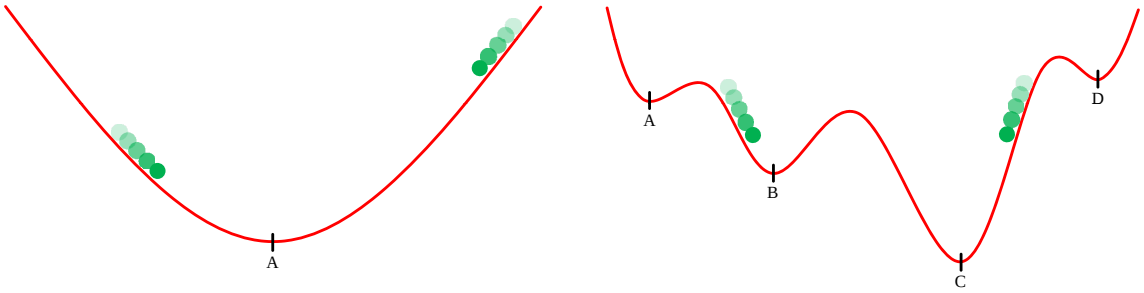
TWO

OPTIMIZATION

After modeling a problem in terms of a random field and posing the corresponding energy minimization problem, we need algorithms which are able to solve such problems. In particular, we are interested in algorithms which allow us to obtain the MAP estimate of discrete random fields. These algorithms belong to the class of discrete or combinatorial optimization techniques. This chapter is dedicated to the introduction of discrete optimization and we will start by an overview of different algorithms which have been proposed and extensively used within the last 30 years. Especially the developments in the last 10 years is mainly responsible for the increasing popularity of discrete optimization. Currently, the most successful algorithms belong either to the class of message-passing methods or to the class of graph-cut methods. We will have a more detailed look at these two classes, while our focus is on the latest developments within the latter class. All our experiments in Chapter 4 on image registration with random fields are based on recent graph-cut based optimization algorithms.

2.1 Energy Minimization

In discrete optimization, the goal is to find a solution to a problem where the solution itself contains integer values, only. This is also sometimes referred to as integer programming [108] and when the solution space is finite we also call this combinatorial optimization [111]. A common approach is to formulate an energy minimization where the minimum of the energy function corresponds to the problem's solution, i.e. the optimal labeling. In the best case, one can express the solution in closed-form and compute it directly. Unfortunately, this usually not possible due to the complexity of the problem. In fact, integer labeling problems are in general NP-hard and we cannot expect to be able to determine the globally optimal solution. There are some special cases for which the optimal solution can be computed in polynomial time; this is the case for binary problems with just two labels or in case the energy fulfills certain conditions and the labels can be linearly ordered. However, for the majority of multi labeling tasks, our hope is on algorithms which can obtain good approximations, i.e. nearly optimal solutions, in a reasonable amount of time.



(a) Convex (b) Non-convex

Figure 2.1: Exemplary convex and non-convex function. The convex energy function in (a) has exactly one global minimum A. This can be easily found independently from the initialization. The non-convex function in (b) has several minima A-D, where only C is the global one. Depending on the initialization, it is more or less difficult to find the global minimum.

2.1.1 The Problem of Non-Convexity

What is it that makes the optimization so difficult? The problem is that in most cases the energy functions we want to minimize are highly non-convex. In Figure 2.1, we illustrate what that means and show the difference between a convex and a non-convex function. While convex functions have one minimum which is also the global one, a characteristic of non-convex functions is that they have many local minima which makes them much more difficult minimize. Consider a gradient-descent optimization which starting at an initial point estimates the gradient of the energy function and moves along the steepest descent. For a strictly convex function it does not matter where the initial point is. The algorithm will always converge to the global minimum (cf. Figure 2.1(a)). This is different for non-convex functions. Depending on where we start the descent, we will end up in different minima (cf. Figure 2.1(b)). One aim of optimization theory is to find algorithms which can overcome this limitation. Discrete optimization methods are currently considered as one of the most powerful techniques for minimizing non-convex functions. Certainly, this one important reason why we consider them for performing image registration, as shown later in Chapter 3.

2.2 History of MRF Optimization

We would like to a give a brief overview of different optimization techniques which are now available for quite a time. Some of them are still quite popular and used in practice – such as simulated annealing or graduated non-convexity – while others – such as iterated conditional modes – have shown to yield inferior results compared with recent advances in discrete optimization and have almost vanished from practical considerations. It is quite interesting to see how these approaches work and how they still influence today's developments. Not all of the presented methods are purely discrete methods. Some were originally applied to problems with continuous variables and make use of variational methods in intermediate steps. We start by introducing simulated annealing which is

the algorithm considered in the first vision MRF paper by Geman and Gemand [36]. The interested reader can find a more comprehensive review of different optimization techniques for random fields in Chapter 9 and 10 in the MRF book [102] by Stan Li.

2.2.1 Simulated Annealing

Random field optimization by simulated annealing (SA) [70, 21] is a general, stochastic method for obtaining global optima for non-convex functions. The idea is based on a physical procedure in which a material is first heated up and then slowly cooled down (annealing) in order to obtain low energy states (or configurations). The temperature is the main parameter of the algorithm and it is iteratively updated according to cooling schedule. Additionally, a certain degree of randomness is introduced which is controlled by the value of the temperature. The randomness makes the algorithm insensitive to the initialization and avoids sticking in local minima. In every iteration, local random changes are made on the labeling. These changes are determined by employing a sampling strategy such as the Metropolis algorithm [106] or a Gibbs sampler [36]. If a local change decreases the energy, it is accepted. If not, it is accepted with a certain probability which depends on the current temperature. While for high temperatures the probability for a change being accepted is also high, the probability gradually decreases with decreasing temperature. The algorithm converges to a low energy configuration, which can be shown to be the global one for certain cooling schedules. Unfortunately, for these schedules the convergence is extremely slow and cannot be used in practice. Faster schedules based on heuristics have to be used such as the one proposed by Geman and Geman in [36] yielding sub-optimal solutions. For more details on simulated annealing we suggest the review by Otten and van Ginneken [109].

2.2.2 Graduated Non-Convexity

Graduated non-convexity (GNC) – proposed by Blake and Zisserman in [11] – is a deterministic optimization strategy which can find good solutions much more efficiently than SA. It belongs to the class of continuation methods. The idea of GNC is to gradually increase the difficulty of the problem during optimization. An optimum of the original non-convex function is obtained by successively optimizing a sequence of functions which are approximations of the original one. The algorithm starts by determining a convex function for which the global optimum can be computed by standard methods such as gradient descent. The solution is then used as initialization for the next round in which an approximating function is optimized which is a little bit closer to the original one. In every round, the approximations are improved until the original function is reached on which the final solution is computed. Exactly how the sequence of intermediate functions is constructed is the key to a successful optimization. In [11], some special cases are presented in which the GNC strategy can obtain the global optimum of the original non-convex function. But except for these cases not much is known about the quality for general functions. The concept of employing functions which are less difficult to optimize is also called *relaxation* where the original, harder function is relaxed and a solution of the

relaxed version is computed first. Another way of relaxation is presented in the following.

2.2.3 Relaxation Labeling

In relaxation labeling (RL) the notion of integer valued label assignments is replaced by so called fuzzy assignments [125]. The state of a random variable is represented by a real-valued vector of size $|L|$ where $|L|$ is the number of labels. The kth entry of the vector reflects the confidence about a label k to be assigned to the random variable. The discrete labeling problem is thus converted into a continuous optimization problem subject to some linear constraints (e.g. the sum of the vector entries has to be one). The solution space for each random variable forms a hyperplane (a simplex) in the multi-dimensional real space $\mathbb{R}^{|L|}$. Relaxation labeling works iteratively and in each iteration the confidence vectors are updated by employing continuous methods. For example, Faugeras and Berthod [28] and Hummel and Zucker [60] use a steepest descent algorithm to compute the updates. In order to obtain a discrete solution after convergence of the RL process, a maximum selection can be used to select the labels with highest confidence. A review on different RL algorithms is given by Kittler and Illingworth [71].

2.2.4 Iterated Conditional Modes

Besag proposed an iterative, deterministic algorithm called iterated conditional modes (ICM) [7] which works with a "greedy" strategy. ICM iterates over the set of random variables and tries to maximize their local conditional probabilities. Each time a variable is visited, the most likely label – the one with the largest increase in probability, respectively largest decrease in energy – is assigned. This assignment is based solely on the local likelihood and prior energy and depends on the currently assigned labels of neighboring variables. ICM converges quite rapidly and can be highly parallelized; variables can be updated simultaneously if they are not neighbors. But in general the obtained solution is only a local minimum of the energy function. The algorithm is extremely sensitive to initialization. In [23] it is proposed to set the initial labeling to the maximum likelihood estimate which results in significantly better performance as reported in [143]. Still, the overall performance is rather weak compared to recent discrete optimization techniques which is also shown in [143].

2.2.5 Highest Confidence First

The last algorithm in our overview is highest confidence first (HCF) [23] which is also a local, deterministic algorithm. HCF tries to improve ICM by introducing a strategy for the order in which random variables are visited. The order is updated in every iteration and depends on a so called *stability measure*. Variables with low stability indicate a higher confidence for changing their labels. These variables are on the top of the list and will be visited at first. Additionally, the label set is augmented by a label named *uncommitted*. This label indicates that a variable has not yet made a commitment to the current configuration. Uncommitted variables are not considered in the energy evaluation for local changes. Experimental results in [23] show an improvement compared to ICM.

2.3 Message Passing

Most of today's MRF optimization problems are either solved by graph-cut methods or by message passing algorithms. The latter ones are discussed in the following, while the principles of graph-cuts are introduced later. The idea in message passing is that a solution to the labeling problem is estimated by iteratively passing local messages around the variables of the random field. Messages are sent from node to node and from the those we can compute what is called *beliefs* about the local configuration. Hence, also the name *belief propagation* (BP) [115].

So, how do the messages look like and how do we compute the beliefs? We come to the exact definition of the messages in a moment – actually there are two different ways how the messages can be generated yielding two different algorithms for BP – but for now let us simply assume that a message m is vector of size $|L|$ (the number of labels). Let us further consider two nodes i and j which are neighbors ($i \in N_j$ and $j \in N_i$) in a first-order random field where the likelihood is encoded on the unary potentials $\psi_i(x_i)$ and the prior is encoded on the pairwise potentials $\psi_{ij}(x_i, x_j)$. Then, by m_{ij} we denote the message which is sent from node i to node j. The $|L|$ entries of the message represent what node i "thinks" about the label assignment of node j. For example, the lth entry of m_{ij} reflects how confident node i is on j getting assigned label l, i.e. $x_j = l$. For convenience, we also write $m_{ij}(x_j)$ to represent this confidence. So, from the viewpoint of node j, every message that j receives from its neighboring nodes $i \in N_j$, tells j what its neighbors think about its labeling x_j. From these messages we can compute the belief $b_j(x_j)$ as

$$b_j(x_j) = \frac{1}{Z_j}\, \rho(x_j) \prod_{i \in N_j} m_{ij}(x_j) \quad . \tag{2.1}$$

Here, the $\rho(x_j)$ is the evidence or compatibility function based on the (unary) likelihood, i.e. $\rho(x_j) \propto \exp(-\psi_j(x_j))$. The constant Z_j is needed for normalization such that $\sum_{x_j \in L} b_j(x_j) = 1$. Intuitively, the beliefs express how probable certain labels are with respect to the likelihood and the confidence of the neighboring nodes. Let us now have a look, how the messages are generated.

2.3.1 Max-Product vs. Sum-Product

There are two possibilities for generating the messages. The first one is used in the so called *max-product* version of BP. The entry $m_{ij}(x_j)$ of a message sent from i to j is computed as

$$m_{ij}(x_j) \leftarrow \max_{x_i \in L} \left(\rho(x_i)\, \rho(x_i, x_j) \prod_{k \in N_i \setminus \{j\}} m_{ki}(x_i) \right) \quad . \tag{2.2}$$

The other version of BP is called *sum-product* and the message update rule is

$$m_{ij}(x_j) \leftarrow \sum_{x_i \in L} \left(\rho(x_i)\, \rho(x_i, x_j) \prod_{k \in N_i \setminus \{j\}} m_{ki}(x_i) \right) \quad . \tag{2.3}$$

Again, $\rho(x_j)$ is based on the likelihood, while $\rho(x_i, x_j)$ is a compatibility function based on the (pairwise) prior, i.e. $\rho(x_i, x_j) \propto \exp(-\psi_{ij}(x_i, x_j))$. Let us understand why this construction make sense. The aim of the message is that node i reveals what it thinks about the labeling x_j. Therefore, node i has to consider three things: (i) assuming node j is assigned x_j, node i has to consider what is the best or most compatible label x_i for itself; this is measured by the compatibility function $\rho(x_i, x_j)$; (ii) but if label x_i is the most compatible one, node i has also to consider what is the likelihood of this label; this is measured by $\rho(x_i)$; (iii) and last but not least, node i needs to consider what the its neighbors think about the label x_i; this is evaluated by considering all incoming messages from neighboring nodes (except of j).

Once every node has sent and received a sufficient amount of messages, based on the beliefs we can compute the configuration of the field. For the max-product BP this is done by evaluating

$$\hat{x}_i = \arg\max_{x_i \in L} b_i(x_i) \quad , \tag{2.4}$$

while the configuration using sum-product BP is obtained by evaluating

$$\hat{x}_i = \sum_{x_i \in L} b_i(x_i)\, x_i \quad . \tag{2.5}$$

Here, the fundamental difference between the two versions of BP becomes clear. In fact, the max-product BP is determining the MAP estimate – the configuration with lowest energy – by maximizing the posterior probabilities. In contrast, the sum-product BP is determining what is called the *minimum mean squared error* (MMSE) estimate by computing the marginal probability distributions. Note, that the MMSE yields configurations which are continuous even if the label set is discrete, since the final labels are computed from a weighted sum of the products between beliefs and labels. Both estimators are used in practice, while in vision applications mainly the max-product BP is used. Some comparisons and experimental results of both versions can be found in [145] for stereo, and in [104] for novelty detection.

2.3.2 Generalizations, Schedules, and Advances

The original BP proposed by Pearl [115] was intended to be used only for graphs without loops, such as Bayesian networks. On tree-like graphs BP works very efficiently and is guaranteed to find the global optimum solution. But nothing actually prohibited the use of BP in graphs with loops. This was the first generalization of BP which is also referred to as Loopy-BP [35]. Despite the fact that in this case the convergence of BP was not guaranteed and little was known about the optimality of the solution, Loopy-BP worked surprisingly well in many vision applications. A lot of work has been done since then to further understand the behavior of BP [160, 34, 161, 154, 89], how it can be generalized to arbitrary graphs [160, 156], and how it can be made more efficiently [30, 95, 87, 117]. An important part of any BP implementation that works on random fields with loops is the message schedule. One can employ serial schedules, where the nodes send their messages one after each other in a specific order, or parallel schedules where all messages are sent

and received, simultaneously. Also some kind of forward-backward strategy is common to ensure that the information is properly propagated over the whole domain (e.g. a 2D image). For example, in [143], a row-column and a scan-line schedule are compared for two variants of BP. A simple illustrative example is given in Figure 2.2(a).

An important advancement of BP is the tree-reweighted message passing (TRW) algorithm by Wainwright [155] and Kolmogorov [83]. TRW decomposes the graph with loops into spanning trees, solves for the labeling of the trees, and reassembles the overall solution. An important feature of TRW is that it simultaneously can compute a lower bound on the energy. The lower bound can be used to assess the quality of the obtained solution, as it is done for the benchmarking in [143]. In fact, the gap between the lower bound and the energy of the obtained solution tells approximately how close we have come to the global optimum. The extended, sequential version TRW-S [83] has guaranteed convergence and was shown to yield excellent results compared to classical BP on common vision applications [143].

2.4 Graph-Cuts

Graph-cuts have become a major tool in discrete optimization, in particular in the vision and imaging community. While in general cuts in graphs can refer to different problems or meanings, here we associate the term graph-cuts with one particular thing: the binary separation of nodes. The origins of graph-cuts as a method for separating nodes into two distinct classes goes back to network theory and is based on the *min-cut/max-flow principle* [32]. Given a graph, the optimal partitioning of nodes can be found by computing a minimum cut [116] – the cut with lowest cost or energy. A cut is represented by a set of weighted edges, which if removed from the graph yields two disconnected sub-graphs. The sum of the weights of removed edges defines the cost of a cut.

An important discovery is that computing the minimum cut is equivalent to solving for maximum flow for which efficient algorithms are available, as shown in [32]. But what is the maximum flow of a graph? We consider a so called *st*-graph which is an augmented version of the original graph where two special nodes are introduced, the source node s and the sink node t (cf. Figure 2.2(b)), which are so called *terminal nodes*. In such a graph the edge weights represent (non-negative) capacities, for example, of imaginary water pipes. Now, the idea of maximum flow is to determine the maximum amount of "water" that can be pushed through the graph from the source to the sink. Of course, this amount depends on the different edge capacities and some edges are getting saturated while successively increasing the flow. Interestingly, the value of the maximum flow is equal to the cost of the minimum cut, and the saturated edges along a path from terminal s to terminal t are the ones belonging to the cut. There are two famous approaches for computing max-flow; the first one is based on *augmenting paths* [32], the second one is based on *push-relabel* [47]. The augmenting paths algorithm is intuitive to illustrate via so called *residual graphs* (cf. Figure 2.2(c)). In each iteration a shortest path from s to t is determined as well as the maximum amount flow that can be pushed along the path. The capacities along the path are decreased and the capacities of reversed edges is increased. When no *st*-path can be found the algorithm stops and the minimum cut can be determined (cf. Figure 2.2(d)).

So far so good, but why is this min-cut/max-flow thing useful for solving discrete labeling problems?

After removing the edges of the minimum cut, every node is either connected to the source or to the sink. Now, if we think of these two special nodes of having some symbolical meaning, for instance, the source corresponds to label "zero" and the sink to label "one", then the node separation corresponds to a binary labeling; every node connected to the source is assigned label "zero", the others are assigned label "one". So, if we find a way to encode the energy function of a binary labeling problem in terms of an *st*-graph min-cut problem – in particular we need to define the capacities of the edges – then we can solve the binary labeling by means of max-flow computation which can be done efficiently in low-order polynomial time [32, 47].

2.4.1 Binary Optimization

Most probably, Greig et al. [48] were the first who made use of the graph-cut approach in computer vision. They tackled the problem of binary image restoration and employed a variant of the max-flow algorithm by Ford and Fulkerson [32]; they also compared their results with simulated annealing and ICM. Their formulation yields an exact solution, i.e. the globally optimal MAP estimate. Boykov and Jolly [15] later used graph-cuts for interactive optimal binary segmentation; they employed an even more efficient max-flow algorithm presented in [16] which is particularly optimized for grid-like graphs often occurring in vision applications. In both works [48, 15], the authors found a way how to set the edge capacities in the *st*-graph such that the min-cut corresponds to the minimum of their problem specific energy function.

The breakthrough of graph-cut based binary optimization is also thanks to the work of Kolmogorov and Zabih [86]. Not only, that they presented a precise characterization of the class of functions that can be minimized, they also derived the necessary condition for these functions to be minimized exactly by graph-cuts. Furthermore, they developed a general purpose *st*-graph construction for this class. From that time on, if one wanted to use graph-cuts for the own binary problem, one could first check if the function is fulfilling the necessary condition, and then simply follow the *st*-graph construction manual; any available max-flow algorithm would then do its job and compute the exact MAP estimate.

Since then, many extensions for graph-cuts have been presented. Just to name a few, for example, Kohli and Torr have developed an efficient dynamic version of graph-cuts [79, 81] and showed how to measure uncertainty in graph-cut solutions [80, 82]. These developments and some more are summarized in the thesis of Kohli [75].

We have briefly mentioned that their exists a necessary condition for functions to be minimized via graph-cuts. Before we come to another quite important advancement which enables us to also minimize functions which do not fulfill this condition, let us discuss what this condition is actually about.

2.4.1.1 Submodularity

Let us consider a binary energy function of the form

$$\mathcal{E}(\mathbf{x}) = \sum_i \psi(x_i) \sum_{ij} \psi_{ij}(x_i, x_j) \quad , \tag{2.6}$$

where $x_i \in \{0, 1\}$. In order to minimize this function via graph-cuts, $\mathcal{E}$ must be *graph-representable* based on the construction rules in [86]. This is specified in the following theorem.

Theorem 2 (Graph-Representable) *A pairwise function $\mathcal{E}$ is graph-representable if and only if each term ψ_{ij} satisfies the inequality*

$$\psi_{ij}(0,0) + \psi_{ij}(1,1) \leq \psi_{ij}(0,1) + \psi_{ij}(1,0) \quad . \tag{2.7}$$

Functions satisfying Equation (2.7) are called *submodular* or *regular* [86]. Note, that unary potentials are always submodular. Intuitively, this condition ensures that the edge weights in an *st*-graph are non-negative. The theorem can be easily extended to potential functions with more than two variables [86].

Many important energy functions fulfill the condition of submodularity. But it happens quite often, that energies, for instance prior terms, which would be more appropriate for our problem at hand are non-submodular. Especially, if we consider that the binary optimization is a crucial component of multi-label optimization. But this is discussed a little bit later. Also we will later see in Chapter 3 that in particular for the task of image registration we are facing non-submodular energies. In the following, we will discuss what we can do in such a case.

2.4.1.2 Minimizing Non-Submodular Energies

A first attempt to deal with non-submodular energies was suggested by Rother et al. [129]. They propose a *truncation* scheme for terms which are non-submodular, which works as follows. Their approach is following the standard construction rules in [86] and whenever a pairwise term is violating Equation (2.7), it is replaced by a truncated term such that the condition holds. This achieved by one of three possible operations; either by decreasing $\psi_{ij}(0,0)$, or by increasing $\psi_{ij}(0,1)$ or $\psi_{ij}(1,0)$ until the inequality is satisfied. For these three operations it is shown that the minimization via graph-cuts is still valid. However, a global optimum of the original function cannot be guaranteed, anymore. Additionally, the number of truncated terms should be small compared to the total number of terms. In that case, graph-cuts with truncation has been shown to yield excellent results in practice [143]. A quite similar but more problem specific scheme was used by Raj and Zabih [120]. They used graph-cuts for image deconvolution and also had to deal with non-submodular terms; these terms are dynamically modified until the submodularity condition is satisfied. A different strategy is presented in [26], where learned statistical priors are encoded on pairwise and triplet terms. Here, the non-submodular terms are simply discarded; but this can decrease the performance as later shown in [85].

A quite different approach for minimizing non-submodular energies is quadratic pseudo-boolean optimization (QPBO) [50, 14, 13]. QPBO can naturally handle energies with both submodular and non-submodular terms. Probably, the first time QPBO has been used for an imaging application is the work of Raj et al. [119] for the task of magnetic-resonance reconstruction. Kolmogorov and Rother [85] provide a detailed description how QPBO can be used for arbitrary non-submodular functions. The basic concept is based on a *reparameterization* of the original energy function into so called *normal form*. Energies of this form can again be minimized by computing a minimum cut on a specially constructed *st*-graph. The construction rules for energies in normal form are also given in [85]. The output of QPBO is a partial labeling $\mathbf{x}$, i.e. $x_i \in \{0, 1, \emptyset\}$ where $\emptyset$ is interpreted as "unknown" or "unlabeled". Before we discuss what this means in practice, let us present two important properties of QPBO:

1. *Persistency*: let $\mathbf{y}$ be a binary labeling and let $\mathbf{z}$ be the "fusion" of the partial labeling $\mathbf{x}$ and $\mathbf{y}$ with $z_i = x_i$ if $x_i \in \{0, 1\}$ and $z_i = y_i$ otherwise. Then, $\mathcal{E}(\mathbf{z}) \leq \mathcal{E}(\mathbf{y})$.

2. *Partial optimality*: there exists a global minimum $\hat{\mathbf{x}}$ such that $x_i = \hat{x}_i$ for all labeled nodes in the partial labeling $\mathbf{x}$.

The partial optimality follows from the persistency, if we take $\mathbf{y}$ as a global minimum. The first property tells us, whatever (initial) labeling we have, if we replace the labels of nodes by the ones that have been labeled in the QPBO result, the energy is guaranteed not to increase. The second property tells us that nodes which are labeled by QPBO are optimal. As pointed out in [128], the usefulness of QPBO in order to minimize non-submodular functions clearly depends on the number of unlabeled nodes. Intuitively, there is some correlation between the number of unlabeled nodes and the number of non-submodular terms. Rother et al. [128] propose several strategies how to deal with unlabeled nodes. It is worth to note, that for submodular energies QPBO yields exactly the same answer as classical graph-cuts, i.e. the globally optimal labeling, with similar efficiency [85].

In summary, we would like to conclude that the introduction of QPBO was an immense progress in graph-cut based optimization which enabled the use of a broader range of energy functions in a principled way. Additionally, as we will see later, QPBO has also lead to a generalization of multi-label optimization and eventually to the point of hybrid discrete-continuous optimization.

2.4.1.3 Higher-Order Potentials

So far, we have mainly discussed pairwise binary energies where the corresponding random field has cliques with a maximum size of two. This is quite natural, if we consider that graph-algorithms such as max-flow estimation work on edges which represent connections between pairs of nodes. However, from Chapter 1 we know that from a modeling point of view there is no restriction on the size of the cliques. In fact, first-order models have a quite limited representational power. Often, it would be desirable to encode richer statistics on potential functions considering more than two variables, i.e. what we call *higher-order potentials*. But how can these energy terms be represented in an *st*-graph construction?

It is a well known fact that any higher-order binary function can be reduced to an equivalent pairwise form [13]. But there are differences how this reduction is performed. Kolmogorov and Zabih [86] already presented *st*-graph construction rules for the case of triplets. This was, at least theoretically, extended to arbitrary higher-order terms by Freedman and Drineas [33]. The employed technique is called *reduction by minimum selection*. Another technique is called *reduction by substitution* which was used by Ali et al. [2]. Recently, Ishikawa [63, 65] pointed out several problems with the latter technique. He proposes a reduction technique also based on the minimum selection for arbitrary higher-order potentials where the energy function is minimized via QPBO graph-cuts. Ishikawa's technique is overcoming some severe restrictions in [33] for functions of even degree. We will make use of this elegant approach later in our framework for image registration. Several works have already shown the advantages of employing higher-order potentials in vision applications [26, 76, 77, 158, 78, 88, 126].

2.4.2 Multi-Label Optimization

Binary labeling is useful for application such as segmentation. But what happens if the label set is $L = \{1, ..., N\}$ with $|L| > 2$, i.e. a multi-label problem? Can we generalize the binary *st*-cut problem to multiple labels?

First approaches in vision were based on *multiway cuts* [17, 9]. In multiway cuts, the graph consists of $|L|$ terminal nodes, instead of just two as in an *st*-graph. The non-terminals, i.e. the nodes corresponding to the random variables, are connected to all terminals. In [17], a "greedy" algorithm is employed in order to compute the cut on this graph which leaves each non-terminal to be connected to exactly one terminal. Little can be said about the optimality or convergence rate of this approach.

A different approach is the graph-construction of Ishikawa [61, 62] which can only be applied if there exists a linear ordering on the set of labels and the prior term has to be convex. The idea is similar to an earlier work by Roy and Cox [130]. Ishikawa constructs a regular *st*-graph with a set of non-terminals of size $|L| \cdot |V|$. So every node $i \in V$ is represented by $|L|$ non-terminals, where each non-terminal defines a specific label assignment. A single minimum cut is then solving the multi-label problem exactly. Which label is assigned to a node depends on where exactly the cut is performed; that is the cutted edge defines the label. This approach was later extended by Veksler to work for truncated convex priors [150]. However, the condition that the labels must be linearly ordered limits the applicability of these approaches.

Today's most powerful multi-label optimization methods are based on a different strategy. Instead of searching for multiway or *st*-cuts on huge graphs, the standard binary min-cut approach is used iteratively yielding very competitive results. This strategy is discussed in the following.

2.4.2.1 Move Algorithms

A powerful algorithm for multi-label optimization based on iterative graph-cuts is α-expansion which has been introduced in the thesis of Olga Veksler [149] and in the work of Boykov et al. [18]. The algorithm is iteratively performing so called *expansion moves*

which are optimally solved by classical *st*-graph cuts. In every expansion move a particular label $\alpha \in L$ is proposed as an alternative for the current labels in $\mathbf{x}$. Every random variable X_i has then to decide whether it keeps its current label x_i or switches to the proposed label α. This is a binary decision which can be solved by min-cut/max-flow on an appropriately constructed *st*-graph. The symbolical meaning of the two terminals becomes "keep" and "change". The decision is made optimally if the underlying energy is submodular. In case of α-expansion and a multi-valued pairwise energy function the necessary condition [86, 85] for submodularity is

$$\psi_{ij}(\alpha, \alpha) + \psi_{ij}(x_i, x_j) \leq \psi_{ij}(\alpha, x_j) + \psi_{ij}(x_i, \alpha) \quad . \tag{2.8}$$

Again, this condition can be easily generalized for higher-order energies [86]. In case of non-submodular functions the expansion move has to be solved by QPBO graph-cuts.

If every label from the label set has been proposed once, this is what we call a *sweep*. After several sweeps, the algorithm converges to a local minimum. However, in case of submodular energies this minimum is guaranteed to be very strong [18]. In many practical scenarios, the minimum in case of non-submodular energies is also very strong, but the optimality cannot be proven.

Following [101], we introduce a move operator $\odot$, such that $\mathbf{x} \leftarrow \mathbf{x} \odot \alpha$ denotes one expansion move. We sketch the α-expansion algorithm in pseudo-code in Algorithm 1.

Algorithm 1: Alpha-Expansion

output: Labeling $\mathbf{x}$

```
1 x ← initializeLabeling() ;
2 for several sweeps do
3 |   foreach α ∈ L do
4 |   |   x ← x ⊙ α ;
5 |   end
6 end
```

Other move algorithms than expansion have been proposed, for instance *swap moves* [149, 18] and *jump moves* [149]. However, the expansion move strategy seems to be the most efficient and accurate one [143].

Komodakis et al. proposed a whole new framework based on principles from linear programming [90, 87, 91, 92, 93]. An important algorithm derived from this framework is FastPD [92]. Similar to α-expansion, FastPD solves a series of min-cut/max-flow problems, but the graph construction is a very different and is based on a primal-dual scheme. In fact, FastPD does not only solve the actual multi-label problem (the primal), but also a dual problem which gives a lower bound on the optimality of the solution. This is similar to TRW-S message passing algorithm [83]. Additionally, FastPD can also be applied to non-submodular energies.

Move algorithms are also sometimes denoted as "very large neighborhood search techniques" [1, 143] which can effectively avoid sticking in local minima of non-convex functions. This is due to the fact, that the set of discrete labels can capture a large (user

defined) search range which allows jumping out of local minima (cf. Figure 2.1). On a side note, recently Serge Belongie, Associate Editor of the IEEE Journal of Pattern Analysis and Machine Intelligence (PAMI), has put the work of Boykov et al. [18] in his all-time favorite top-ten list of PAMI articles with the reason "for a few years there it seemed that every problem I was working on could be written down with a cost function that Yuri, Olga and Ramin's code could solve for me quickly and accurately". We should note that back then when Boykov et al. published their work on expansion moves the applicability of their method was mainly limited to submodular energies. Today's advancements on minimizing non-submodular energies such as QPBO or FastPD allow tackling an even broader range of problems. In particular, the boundaries between discrete and continuous optimization are beginning to blur, as we will see in the following.

2.4.2.2 Discrete-Continuous Optimization

A recent, very promising advancement is the concept of *fusion move* optimization [100, 101]. Fusion moves are based on QPBO graph cuts and generalize the move operator $\odot$. Let us consider any two (sub-optimal) labelings $\mathbf{x}^1$ and $\mathbf{x}^2$ of a problem defined by energy $\mathcal{E}(\mathbf{x})$. Now, the idea of fusion is to combine (or fuse) these two labelings into a new one $\mathbf{x}^f$, such that $\mathcal{E}(\mathbf{x}^f) \leq \mathcal{E}(\mathbf{x}^1)$ and $\mathcal{E}(\mathbf{x}^f) \leq \mathcal{E}(\mathbf{x}^2)$. We can express all possible combinations by an auxiliary binary labeling $\mathbf{y}$ as

$$\mathbf{x}^f \leftarrow \mathbf{x}^1 \bullet (1 - \mathbf{y}) + \mathbf{x}^2 \bullet \mathbf{y} \quad , \tag{2.9}$$

where $\bullet$ is the element-wise product. So, all we need to compute is the optimal binary labeling $\mathbf{y}$, and this is done by a QPBO graph-cut. Note that of course classical *st*-cuts could be employed as well if the (pairwise) energy satisfies the following necessary condition of submodularity

$$\psi_{ij}(x_i^1, x_j^1) + \psi_{ij}(x_i^2, x_j^2) \leq \psi_{ij}(x_i^1, x_j^2) + \psi_{ij}(x_i^2, x_j^1) \quad . \tag{2.10}$$

However, this is less likely than the condition for α-expansion, since here, the labelings $\mathbf{x}^1$ and $\mathbf{x}^2$ could theoretically be anything, for instance random. QPBO seems to be a natural choice for the binary optimization. The combination of two labelings is called a fusion move; we use the move operator to indicate this as

$$\mathbf{x}^f \leftarrow \mathbf{x}^1 \odot \mathbf{x}^2 \quad . \tag{2.11}$$

Now, we also see why fusion moves are a generalization of previous move strategies. For instance, we could set one the two labelings to the constant labeling $\mathbf{x}^\alpha$ where each node is labeled with α; the fusion move becomes an expansion move.

The labelings which are going to be fused are sometimes also called *proposals*. And since these proposals can be more or less anything, fusion moves also allow some kind of *discrete-continuous* optimization. For instance, the proposals could be solutions generated by two completely different algorithms, where maybe one is using α-expansion and the other one belief propagation. The two solutions can be fused, and the energy is guaranteed not to increase. But the proposals could also be continuous where the values are not

necessarily within a predefined discrete label set. We could fuse solutions from continuous optimization methods and the result will also be continuous. This approach has been successfully used for stereo matching [158, 157] and optical flow [99]. Recently, Ishikawa [64] proposed to use gradient-based proposals where the proposal itself is computed from the gradient of the energy function such that we can perform a gradient-descent via fusion moves. To summarize, we believe that there is great potential in this new type of optimization and space for creativity on how to generate the best proposals.

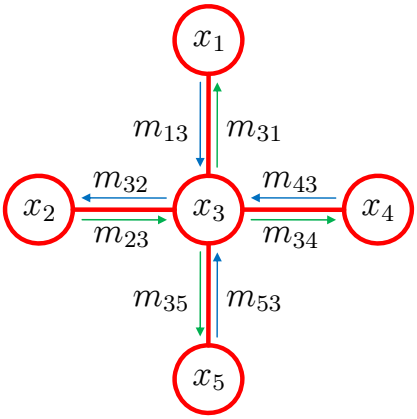
2.5 Message Passing vs. Graph-Cuts

We would like to conclude this chapter on optimization by pointing out some works on comparisons between message passing and graph-cuts. A lot of different experiments have been carried out to investigate which of the two approaches works better in case of discrete multi-labeling. In order to come to the point, there is no clear answer. It highly depends on the specific application, type of energy, graph topology, and others. Already some time ago, Tappen and Freeman [145] compared swap moves [18] with two variants of belief propagation, a max-product and a sum-product version, for stereo matching with identical MRF parameters. They concluded that the results were comparable while graph-cuts usually resulted in lower energy solutions. Mahamud [104] compared an (unoptimized) version of graph-cuts again with max-product and sum-product BP for novelty detection, which is a binary problem. He found that BP gives slightly better results. Kolmogorov and Rother [84] compared α-expansion [18] with TRW-S [83] and a max-product BP on stereo matching with highly-connected graphs. They found that α-expansion clearly outperforms the two message passing approaches.

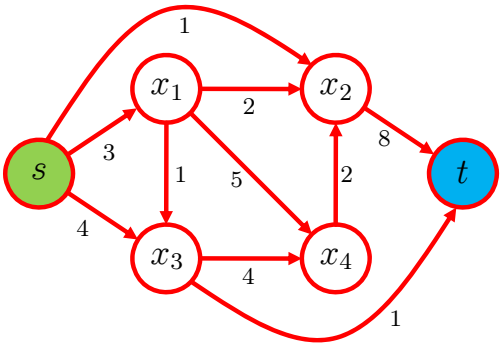
The probably most comprehensive study so far was conducted by Szeliski et al. [143][1]. They investigated four different discrete labeling problems: binary segmentation, stereo matching, photomontage, and restoration (or denoising). The tested algorithms were TRW-S, α-expansion, swap moves, and different variants of max-product BP. There was no clear overall winner, and the algorithms performed quite differently on the different problems. For the two graph-cut based methods, one could conclude that expansion moves performed always better or comparably well as swap moves. TRW-S was sometimes able to find the (almost) globally optimal solution when graph-cuts did not perform well at the same time. On other sequences, graph-cuts, and in particular, α-expansion performed best. In general, the graph-cut methods were the most efficient ones with respect to computation time.

There is definitely further need for conducting such experimental studies. In particular, in future comparisons QPBO based fusion moves or the FastPD method should be included. There is also an increasing interest in higher-order random fields and previous studies have only considered pairwise MRFs. In general, one should say that it is worth trying out different optimization methods for the problem at hand. Especially, since many authors made their own implementations available for download and public use.

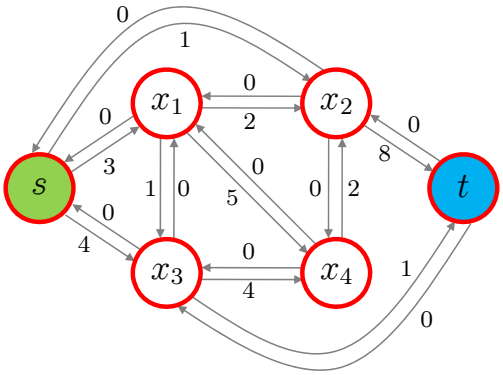
[1] `http://vision.middlebury.edu/MRF/`



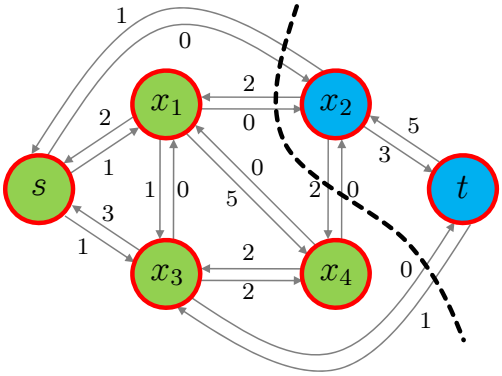
(a) Message passing



(b) *st*-graph



(c) Initial residual graph



(d) Minimum cut

Figure 2.2: Message passing and graph-cuts. A simple example of message passing is shown in (a). The green arrows indicate the direction of messages in a forward pass. The blue arrows indicate the backward pass. An exemplary *st*-graph is shown in (b). The corresponding initial residual graph is shown in (c). After max-flow computation, saturated edges form s to t are cut. These edges constitute the minimum cut which separates the set of nodes in two distinct classes.

CHAPTER

THREE

IMAGE REGISTRATION

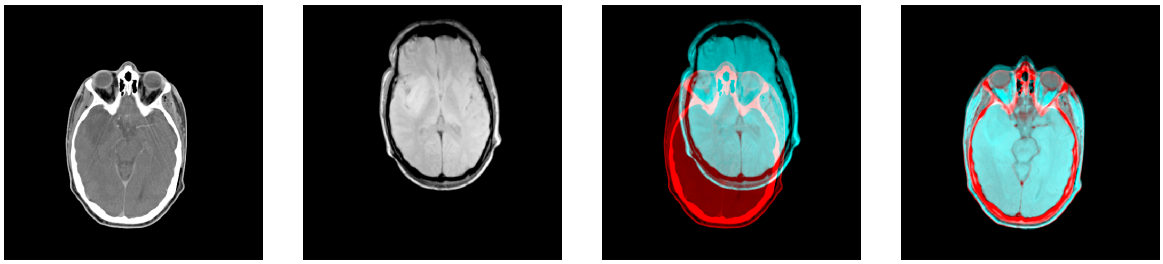
In this chapter we present our approach for image registration based on random fields. The theory of random fields has been earlier introduced in Chapter 1. Here, we build upon this theory and derive a general framework applicable to various registration scenarios. We start with a brief introduction to the exciting field of registration and see why this is a challenging problem worth to be considered from a novel perspective. Then we come to our contribution which is the introduction of discrete random fields for tackling linear and non-linear registration. In order to provide an intuitive understanding, we will mainly present the general concept and the derivation of the different methods, while leaving out some implementation specific details to avoid distraction. These details, such as the actual setting of certain parameters, can be found in our papers referenced throughout the text. At the end of this chapter some prior work is reviewed. We conclude by a discussion and an outlook on open questions. Different experiments and applications in which our methods have been shown to be very efficient are later presented in Chapter 4.

3.1 Introduction

We would like to start by presenting some basics and core components involved in any registration method. But let us first clarify, what we mean by "image registration":

Informal Definition 1 (Image Registration) *Image registration is the task where one seeks spatial transformations which align a set of images.*

We further say that images are aligned when corresponding structures overlap after applying the transformations and superimposing the images. The transformation determines a relationship between the coordinate systems of the images. In the following, we focus on *pairwise registration* where we seek one transformation mapping the first image to the second. It is common to assume that the second image remains fixed – we call this the *target image* – while the first image is the one undergoing a transformation – we call this the *source image*. In literature, the target image is also sometimes referred to as the reference or fixed image, and the source image is also denoted as the floating or moving



(a) CT (b) MRI (c) Initial (d) Registered

Figure 3.1: CT-MRI brain registration. The original images are shown in (a) and (b). The initial pseudo-color overlay in (c) shows the misalignment. In (d) the corresponding structures overlap after performing registration.

image. The images are considered to have the same dimensionality, which is either 2D or 3D.

3.1.1 Why do we need Registration?

Image registration enables us to depict correspondences between images. But why is this useful? The need for registration is ubiquitous and often a key component in many imaging and computer vision applications. The medical domain is one of the most prominent areas. In computer-aided diagnosis, the fusion of images acquired with different sensors can support the assessment of pathologies. Often, the combination of computed tomography (CT) and magnetic resonance imaging (MRI) provides a broader picture of the patient's anatomy to the physician. When exploring properties such as shape, appearance or size of a specific area (e.g. an organ), it might be beneficial to have the information from different scans available. For example, CT and MRI are based on completely different physical principles. While CT images often have high-resolution and bony and calcified structures are clearly visible, MRI can provide higher contrast in soft-tissue areas. Here, we need registration to establish correspondences between the two modalities (see Figure 3.1). Similarly, the fusion of functional and anatomical images is of great interest. Once registered, the functional image can tell us something about *what* is happening, while the anatomical image tells us *where* it is happening. Other examples from the medical domain where registration is important are follow-up studies, computer-assisted navigation, or the creation of anatomical atlases. For an overview, we refer to the upcoming book by Paragios et al. [112].

Also many non-medical applications can benefit from registration. For instance, recovering the apparent motion of objects from a video sequence is known as the optical flow problem. By determining the correspondences between individual image points, we can compute a motion field. This information is then further processed in applications such as surveillance, motion tracking, and robotics [58, 113]. Other applications such as image stitching and generation of panoramic views [142] also require registration between camera pictures.

We will discuss some applications – from the medical and non-medical domain – in

more detail later in Chapter 4. Note, there exist many synonyms for the task of image registration such as image alignment, image matching, motion estimation, or image correspondence problem. More detailed discussions on the huge field of registration can be found in the books by Hajnal et al. [49] and Modersitzki [107], in the surveys of Maintz and Viergever [105], Zitova and Flusser [166], and in the tutorial by Richard Szeliski [142].

3.1.2 How do we do Registration?

There are many ways how we can establish registration of two images and likewise the amount of literature on this topic and the number of different approaches and algorithms is numerous. The reader may forgive that we are not providing a comprehensive review of methods but rather present the basic ideas on which many approaches have been build on. Later, we will discuss some related work and classical approaches considered for this task. For more details, we refer the interested reader to the above mentioned surveys and the references therein.

Most probably the simplest way to perform registration (at least from an engineering point of view) is to ask someone – most probably an expert, e.g. a radiologist – to manually select a sufficient number of corresponding points in the two images to be aligned. From these points, we can easily compute a transformation and the registration is done [3, 12, 148, 31]. Despite its simplicity such an approach is not very effective and often doomed in practice for several reasons. The manual selection of corresponding points can become a very tedious task. Often, even for an expert, it is not easy to identify correspondences, in particular when the images are acquired with different modalities. Also, the manual interaction can be very time consuming. Depending on the application and the considered type of transformation, a *sufficient* number of points can mean that several hundreds or even thousands of correspondences are necessary to determine a proper alignment. A solution to this dilemma is the employment of methods for automatic detection of point correspondences. The basic idea is to extract a set of candidate points with a certain special appearance or local structure. Then, a robust algorithm tries to detect corresponding points between the candidates and removes outliers at the same time. Finally, the transformation is computed from the detected correspondences. These methods belong to the class of feature-based or landmark-based registration [8, 27, 68, 124, 138, 24]. The advantage of feature-based registration is the efficiency from a computational point of view. However, since the transformation is computed solely from a rather sparse set of feature points there is no guarantee on the quality of the alignment in areas where no features have been extracted.

A quite different approach is intensity-based registration. Here, the registration is performed by considering directly and only the image intensities without the need of extracting any features or landmarks. The basic idea is to evaluate the quality of an image alignment by looking at the intensities of overlapping image points. When images are perfectly aligned the intensities of corresponding points should be somehow similar. An advantage of intensity-based registration is that all image points can contribute to the estimation of the transformation, and not only a sparse set of feature points whose identification can be a critical source of errors. However, the computational complexity is

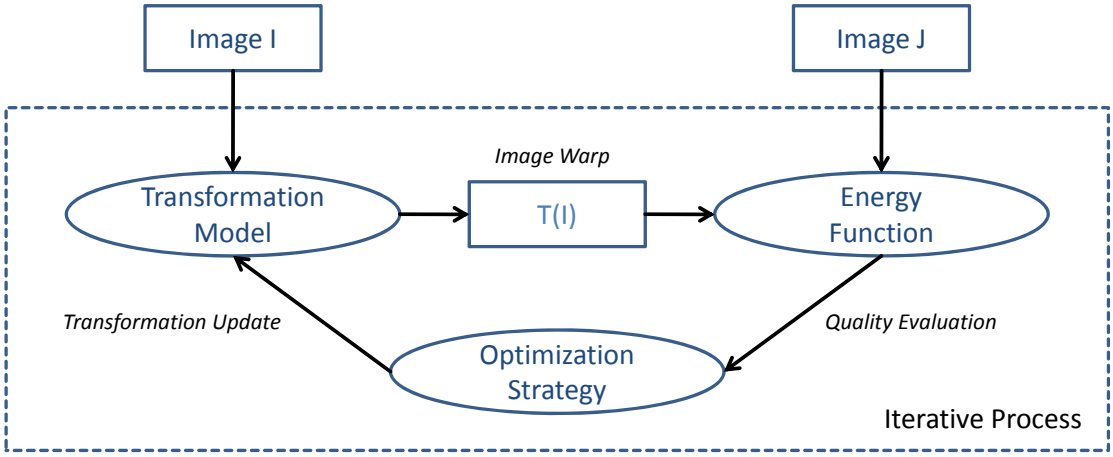


Figure 3.2: Schematic illustration of the iterative registration process.

often much higher compared to feature-based registration and the images should exhibit a sufficient initial overlap.

In practice, these two approaches are sometimes combined. For example, the feature-based part can be used for the initial pre-alignment and the intensity-based part for the refinement [66]. Other hybrid methods [29, 20, 55, 110, 140] exist which make use of feature points and intensities, simultaneously. Our MRF registration framework mainly belongs to the class of intensity-based registration but is not restricted to this type of registration. Possibilities how information from landmarks or feature points can be (additionally) embedded in our framework are presented in [41, 140]. Let us now have a closer view on the basics of intensity-based registration and its core components.

3.2 Intensity-Based Registration

A common approach for intensity-based registration is the formulation as an energy minimization problem. Let us consider the registration of two images I and J where I is the source image undergoing a transformation denoted by $I' = I \circ T$; the image J is the target. The optimal transformation $\hat{T}$ which aligns the two images can be obtained through the following minimization:

$$\hat{T} = \arg\min_{T} \mathcal{E}(I \circ T, J) \quad . \tag{3.1}$$

From a probabilistic point of view this is equivalent to maximizing the posterior probability $\rho(T|I, J)$ (cf. Section 1.4 in Chapter 1). The energy function $\mathcal{E}$ is based on a similarity measure[1] which evaluates how similar the images are with respect to the transformation T. This function is in general highly non-convex and a direct solution to Equation (3.1) is

[1]The term similarity measure is sometimes misleading when we talk about a minimization problem. Here, we always assume that a lower energy indicates that the images are more similar and better aligned.

difficult to compute or not possible. To this end, the minimization is performed through an iterative process where in each iteration one tries to find an optimal update $\Delta\hat{T}$. Each update should minimize the energy a little bit further until a minimum of $\mathcal{E}$ is obtained. Such a process is sketched in pseudo-code in Algorithm 2.

Algorithm 2: Intensity-based Image Registration

input : Images I, J
output: Transformation T

1 $T \leftarrow$ `initializeTransformation()` ;
2 **repeat**
3 | $\Delta T \leftarrow$ `computeUpdate(`I`,`J`,`T`)` ;
4 | prevEnergy $\leftarrow \mathcal{E}(I \circ T, J)$;
5 | newEnergy $\leftarrow \mathcal{E}(I \circ (T + \Delta T), J)$;
6 | **if** newEnergy $<$ prevEnergy **then**
7 | | $T \leftarrow T + \Delta T$;
8 | **end**
9 **until** *convergence*;

The convergence criterion can be for instance based on the change of energy. If there is only a very small difference between the previous energy and the new energy, we might stop the algorithm. Another possibility is to set a fixed number of iterations. Note that the update is only added to the current transformation if the energy decreases.

Line 3 in which the update is computed is the most critical part of the whole algorithm. The performance of the registration highly depends on the method which is used in this step. Ideally, the updates should yield an efficient energy minimization in terms of speed and the obtained minimum should be a strong one, in best case a global one. Additionally, the quality of the updates should be independent of the actual energy function, which means we would like to have an algorithm which yields excellent results in any registration scenario. By introducing random field optimization as an approach for computing these updates, we aim at providing both computational efficiency and strong solutions.

In general, intensity-based registration involves the selection of three main components: (i) a transformation model, (ii) an energy function, and (iii) an optimization strategy. In Figure 3.2 we show a schematic illustration of the registration process and the role of the three components. An important part of the energy function is the intensity-based similarity measure. In the following, we present some popular choices. A brief overview of different transformation models is given afterwards and the details are discussed together with the optimization strategy when we come to our registration framework.

We should also note that intensity-based registration is commonly solved with multi-resolution and scale-space strategies. For instance, in all our actual implementations we make use of Gaussian image pyramids. The idea is to start the registration on smoothed and downsampled versions of the images and subsequently increase the resolution. This strategy can help to reduce the computation time but also to avoid sticking to local minima in the beginning.

3.2.1 Similarity Measures

A variety of similarity measures can been considered for intensity-based registration. Which of the measures is employed in practice highly depends on the application and in particular on the modalities of the images, i.e. the way the images have been acquired. The selection should always reflect the assumptions made on the intensity relationship between the source and target image. For instance, if we can safely assume that the intensity distributions of the two images are the same and if corresponding points exhibit the same intensity values, simple measures based on intensity differences are sufficient. If the two images are acquired with different modalities, we have to consider more complex measures which for instance determine a statistical relationship of intensities. Similarity measures can be separated into two categories, pointwise measures and region-based measures. We will introduce the most popular measures for both categories in the following. Note that we are consistently using the term similarity measure for all of them, despite the fact that some criteria are actually dissimilarity measures. However, every criterion can be converted such that it fulfills our assumption that lower energies correspond to higher similarity and better alignment. Here, we can ensure this by simple negation or inversion.

3.2.1.1 Difference Measures

The simplest similarity measures are based on the differences of intensities. Here, we consider the (normalized) sum of absolute differences (SAD) which is defined as

$$\mathcal{S}_{\mathrm{SAD}}(I, J) = \frac{1}{|\Omega|} \sum_{p \in \Omega} |I(p) - J(p)| \quad , \tag{3.2}$$

and the (normalized) sum of squared differences (SSD)

$$\mathcal{S}_{\mathrm{SSD}}(I, J) = \frac{1}{|\Omega|} \sum_{p \in \Omega} (I(p) - J(p))^2 \quad . \tag{3.3}$$

The normalization by the number of points make the measures independent of the size of the overlap domain Ω. The SAD criterion is slightly more robust to outliers while the SSD criterion tends to over-penalize outliers. Both measures are belonging to the class of pointwise measures. They are extremely easy to implement and very efficient to compute. Intensity differences are only considered in mono-modal registration where the images are acquired with the same modality.

3.2.1.2 Statistical Measures

More sophisticated similarity measures are based on image statistics [167]. They belong to the class of region-based measures since they extract additional information from the image domain Ω. The advantage of statistical measures is that corresponding image points do not necessarily have to have the same intensities. This is actually very often the case because of illumination changes, inhomogeneities, noise, and differing acquisition

processes. A measure which makes less strict assumptions on the two intensity distributions than difference measures is the correlation coefficient (CC). The CC criterion is defined as

$$\mathcal{S}_{\mathrm{CC}}(I,J) = \frac{\sum_{p\in\Omega}(I(p)-\mu_I)(J(p)-\mu_J)}{\sqrt{\sum_{p\in\Omega}(I(p)-\mu_I)^2}\sqrt{\sum_{p\in\Omega}(J(p)-\mu_J)^2}} = \frac{\mathrm{cov}(I,J)}{\sigma_I\sigma_J} \quad , \tag{3.4}$$

where μ_I and μ_J are the two means and σ_I and σ_J the standard deviations of the image intensity distributions. The CC takes values from $[-1,1]$, where 1 indicates a perfect linear relationship, 0 indicates no linear relationship, and -1 an inverse linear relationship. By simply taking $1-\mathcal{S}_{\mathrm{CC}}$ we can map the values on the range $[0,2]$ in order to make CC suitable for minimization. The CC criterion makes less strict assumptions on the intensity relationship than difference measures. Still, it is assumed that a linear relationship exists. A measure which goes beyond this restrictions by assuming a pure statistical relationship is based on mutual information (MI) [103, 152, 141] and defined as

$$\mathcal{S}_{\mathrm{MI}}(I,J) = \mathcal{H}(I) + \mathcal{H}(J) - \mathcal{H}(I,J) \quad . \tag{3.5}$$

Here, $\mathcal{H}(I)$ and $\mathcal{H}(J)$ are the marginal entropies, and $\mathcal{H}(I,J)$ is the joint entropy of images I and J. The entropies are defined as

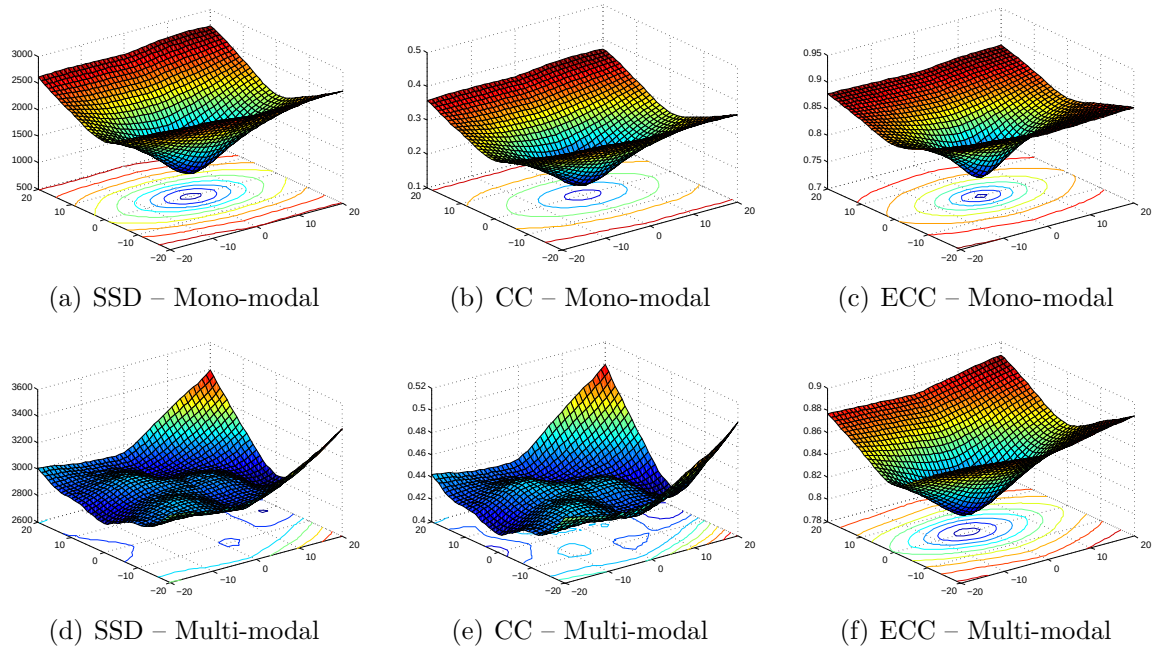
$$\mathcal{H}(I) = -\sum_i \rho(i)\log\left(\rho(i)\right) \quad \text{and} \quad \mathcal{H}(I,J) = -\sum_{i,j}\rho(i,j)\log\left(\rho(i,j)\right) \quad , \tag{3.6}$$

where $\rho(i)$ and $\rho(i,j)$ are the marginal and joint intensity distributions. So, using MI as a similarity measure needs some more computation steps than the simple difference measures or the CC criterion. Given two images, we first estimate the joint intensity distribution. There are several possibilities how to do this [121]; a simple and straightforward way is by means of a discrete joint intensity histogram with a fixed number of bins. From the (normalized) joint histogram we can read out the marginal and joint probabilities by iterating over the bins of intensities. A popular variation and normalized version of MI is the entropy correlation coefficient (ECC) [103] taking values from $[0,1]$. The ECC is defined as

$$\mathcal{S}_{\mathrm{ECC}}(I,J) = 2 - \frac{2\,\mathcal{H}(I,J)}{\mathcal{H}(I)+\mathcal{H}(J)} \quad . \tag{3.7}$$

Again, taking $1-\mathcal{S}_{\mathrm{ECC}}$ makes it suitable for minimization. Such purely statistical measures can be used for multi-modal registration where no linear or even no non-linear relationship between the image intensities of corresponding points exist. The classical application of MI-based measures is for instance CT-MRI registration (cf. Figure 3.1).

Other popular statistical measures are for instance the correlation ratio [122], Kullback-Leibler divergence [25], or Jensen-Renyi divergence [54]. Recently, quite successful attempts have been made in order to learn the statistical relationship between images [96, 114, 19].



(a) SSD – Mono-modal (b) CC – Mono-modal (c) ECC – Mono-modal

(d) SSD – Multi-modal (e) CC – Multi-modal (f) ECC – Multi-modal

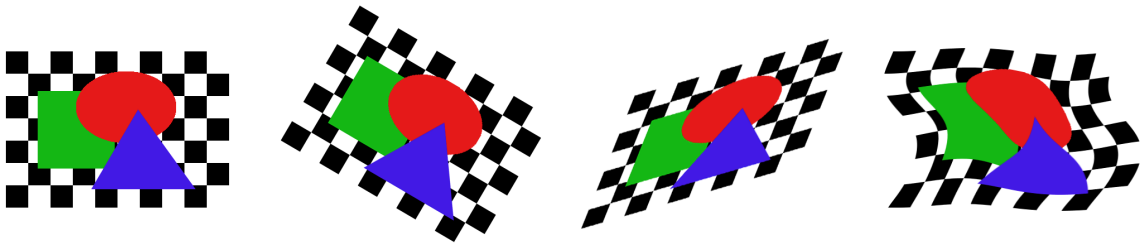
Figure 3.3: Behavior of similarity measures in mono- and multi-modal scenarios.

3.2.1.3 Behavior of Similarity Measures

In Figure 3.3 we show the results of a small experiment which should illustrate the behavior of the different similarity measures. We took a pair of corresponding 2D images from registered CT and MRI scans (cf. Figure 3.1); we add Gaussian white noise to the images. In order to simulate a multi-modal registration, which allows us to investigate the similarity measures, we used the MRI slice as source and the CT slice as target image. We translated the source image within an range of ± 20 pixels along the x- and y-axis; at each location we evaluated the SSD, the CC, and the ECC. Then we did the same in a simulated mono-modal scenario. Here, we used the MRI for both the source and the target image. The perfect alignment is of course when non translation is applied to the source image, i.e. at point $(0, 0)$, since the images are already registered. We clearly see that all three measures have a well defined minimum at the correct location in case of mono-modality. For the multi-modal case, only the ECC measure is able to correctly determine the right translation.

3.2.2 Transformation Models

Another crucial component of image registration is the transformation model. The type of transformation defines which motions or movements the source image can undergo. Also the selection of the transformation model highly depends on the application and should be based on the motions that we expect in the images. Linear transformations [53] are suitable when translation, rotation, scaling or shearing is sufficient to align the images (cf. Figure 3.4(b) and 3.4(c)). For example, if only the pose and orientation of the imaging device has been changed from one image to another, linear transformations are sufficient.



(a) Identity (b) Rigid (c) Affine (d) Non-linear

Figure 3.4: Different types of image transformations. Transformations (a)-(c) are linear mappings, while (d) is non-linear.

In medical applications, linear transformations are considered when rigid structures, such as the head of a patient or bony structures, are to be registered. If objects might have changed their shape, for instance when the images are acquired at different breathing phases, we need transformations which are more flexible and able to recover deformations (cf. Figure 3.4(d)). Similarly, when multiple objects with different motions are present within one scene, non-linear transformations [57] need to be employed.

Registration with linear transformation models is sometimes referred to as *rigid*, *affine* or *global* registration. While registration with non-linear transformations is also called *deformable*, *non-rigid*, or *local* registration. Figure 3.4 shows an overview of different types of transformations. Most transformations can be represented by a finite set of parameters. The number of parameters is closely related to the *degrees of freedom* (DOF) of a transformation. Basically, all previously used transformation models for image registration fulfill the assumption of parameterization if we follow the viewpoint of [165]. This is an important property when considering registration with random fields.

3.3 Registration with Random Fields

We have seen different possibilities for the similarity measure which is one part of the energy function and we roughly discussed the role of the transformation model. The last component of intensity-based registration is the optimization strategy, and in particular the part of computing the updates ΔT of the transformation (cf. Equation (3.1) and Algorithm 2). Here, random fields come into the game. We propose discrete random field optimization as a powerful alternative for this component. The general concept of random fields and discrete optimization have been discussed in the previous two chapters. Now, we demonstrate how this concept can be applied to image registration.

Let us formulate the registration problem in MRF terminology. The set of images to be registered is the given data or the observation. So what is the role of the random variables? We assume that the sought transformation T can be represented by a finite set of n parameters $\Phi = \{\varphi_1, ..., \varphi_n\}$; we also write T_Φ to emphasize the parameterization of the transformation. Then, in every iteration the task is to determine the optimal update $\Delta\Phi$. If we associate the update of a single parameter $\Delta\varphi_i$ with a hidden variable X_i of a

random field $\mathbf{X} = \{X_1, ..., X_n\}$, then every labeling $\mathbf{x}$ corresponds to an update $\Delta\Phi$ and accordingly to a new transformation $T_{\Phi+\Delta\Phi}$. So, the task of registration is converted into successive labeling problems where we have to solve one labeling problem per iteration.

In Chapter 1 we have introduced a general recipe for posing an optimization problem as a labeling problem. So far, we have identified the role of the random variables; they represent the updates of transformation parameters. Now, the challenging task is to explicitly define the posterior energy of the random field such that it matches the registration energy. To this end, we have to define a set of cliques and their associated potential functions. Furthermore, we have to discuss how to define the set of labels. But all this depends on the actual type of transformation. Let us start with the case where the transformation is non-linear. Afterwards, we introduce our random field model for the linear case.

3.4 Non-Linear Registration

In non-linear registration we are trying to recover a non-linear transformation, hence the naming. In Section 3.2.2 we have mentioned that this type of transformation is necessary to recover the deformation of an image, or to be more precise, the deformation of objects visible in an image. But how does such a transformation look like? It is convenient to define it as a decomposition into two terms, an identity transformation Id and a dense displacement field D, such that we get $T = Id + D$. A displacement is simply a vector describing the motion of an image point p. The new location of p is then defined as

$$T(p) = p + D(p) \quad . \tag{3.8}$$

If we insert this transformation model into the very general registration Equation (3.1) and employ a suitable similarity measure $\mathcal{S}$, we could perform non-linear registration by solving the minimization problem

$$\hat{D} = \arg\min_{D} \mathcal{S}(I \circ (Id + D), J) \tag{3.9}$$

It is clear that the actual task of non-linear registration is to recover the displacement field D. This turns out to be a quite challenging task: (i) assuming the registration energy is solely based on similarity measure $\mathcal{S}$; recovering a higher-dimensional displacement field from a one-dimensional intensity signal is an ill-posed problem [147]; there is little hope that the optimization yields reasonable registration results; (ii) the number of unknowns is often enormous; if we consider three-dimensional registration of moderately sized images with 256^3 voxels, then the corresponding displacement field has more than 16.7 millions of parameters, each corresponding to a three-dimensional displacement. In total we would have more than 50 million degrees of freedom.

Before we define our random field, we have to find solutions to deal with both issues. Let us first present a solution to the first issue which is more of theoretical nature but has a dramatic impact in practice.

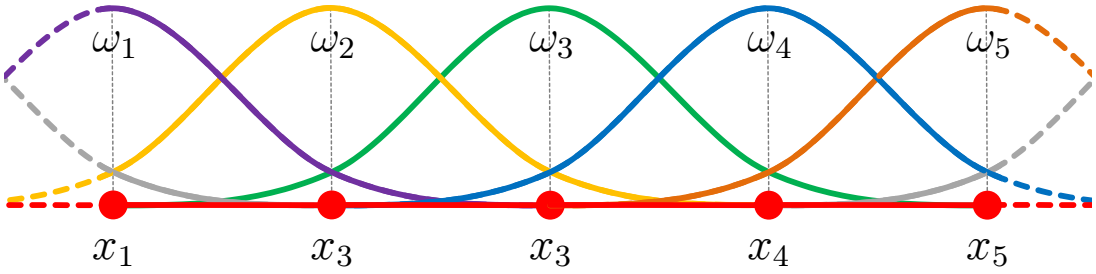


Figure 3.5: Dimensionality reduction. The dense domain (red line) is represented by a sparse set of control points (red dots). Weighting functions ω control the influence of the control points.

3.4.1 The Need for Regularization

One popular way to overcome the ill-posed formulation is to add a second term to the energy function which renders the optimization problem well-posed

$$\hat{D} = \arg\min_{D} \mathcal{S}(I \circ (Id + D), J) + \mathcal{R}(D) \quad . \tag{3.10}$$

The function $\mathcal{R}$ is also called a *regularization term*. Regularization plays an important role in many optimization problems. As the name suggests, it allows us to regularize or constrain the space of solutions, for instance by penalizing undesired or unrealistic displacements fields. We could define $\mathcal{R}$ such that it returns high energy values if D contains high gradients which would indicate non-smooth deformations. In fact, very often smoothness is a desired property of deformations. If $\mathcal{R}$ is favoring smooth displacement fields, we also call it a *smoothness term*. Other wanted properties could be that the displacement field should be invertible and free of foldings and other singularities.

On a side note, we actually have seen the concept of regularization much earlier, namely in Chapter 1 in the three vision examples. The prior energy of an MRF is often nothing else than a regularization term defined on the space of labelings. For instance, an assumption that neighboring pixels should have similar labels is a simple smoothness condition. Let us now present a solution to the second issue, the enormous amount of parameters.

3.4.2 Dimensionality Reduction

Intuitively, we could introduce a random variable for every image point p. However, such an approach would result in an extremely huge field for which optimization is very inefficient, time consuming and memory demanding. We need to reduce the dimensionality of the problem. An elegant way of doing this is by reparameterization of the dense displacement field. Let us consider a set of k control points distributed along the image domain. Furthermore, let k be much smaller than the number of image points n. The dense displacement field can be defined as a linear combination of control point displacements

$\Phi = \{\varphi_1, ..., \varphi_k\}$ with $\varphi_i \in \mathbb{R}^d$

$$D(p) = \sum_i^k \omega_i(p)\, \varphi_i \quad . \tag{3.11}$$

Here, ω_i corresponds to an interpolation or weighting function (cf. Figure 3.5). It determines the influence of a control point i to the point p. Commonly, the closer a point p is located to i the higher the influence of the displacement φ_i. So, the actual displacement of an image point is now computed by weighted sum. The important thing is, that in order to obtain a dense displacement field, and thus a deformation of an image, we just have to manipulate a few control points. For convenience, we denote the displacement of point p by $\delta_p \equiv D(p)$.

A popular concept is based on free-form deformations (FFDs) [135] with cubic B-splines as weighting functions. In the common FFD approach, the control points are uniformly distributed over the image domain. Each control point has only local support on the displacement field, and thanks to the cubic basis functions the resulting deformation is guaranteed to be C^2 continuous (cf. Figure 3.6). FFDs are also very efficient to compute, since a lot of things can be pre-computed due to the regular structure. However, the lattice-like topology of FFDs has some drawbacks. The control points do not adjust well to the visible structures. This might be problematic if multiple objects undergo different kind of motions. A regular grid will always have difficulties at the motion boundaries. Here, transformation models with arbitrarily distributed control points, e.g. based on triangulations, are beneficial. Those can be adjusted to the actual image content, however usually their evaluation comes at higher computational costs.

Now we have seen possible solutions for the two challenges or issues mentioned earlier. Based on this, we can now derive our models for non-linear registration. Here, we propose two alternatives. The first model yields an extremely efficient registration method and is based on free-form deformations and first-order Markov random fields [43, 44]. The second one is based on triangulations and higher-order conditional random fields [38]; for the latter one we make use of the most recent advances in discrete optimization.

3.4.3 The First-Order MRF Model

For our first-order model we employ free form deformations as the transformation model. The first time we have presented this model was in [43] and later more comprehensively in [44]. The FFD control points are distributed on a regular grid and the updates on their displacements $\Delta\Phi$ are associated with the random variables $\mathbf{X}$. So a label assignment $X_i = x_i$ is equivalent to adding a displacement $\Delta\varphi_i$ to the current position of control point i. From now on, we treat labels directly as displacements, i.e. $x_i \equiv \Delta\varphi_i$. Similar to the vision examples in Chapter 1, we employ a simple neighborhood system for the random variables which follows the natural layout of the regular FFD grid, i.e. a 4-connected neighborhood in case of 2D and 6-connected one in 3D. The first-order MRF energy is defined as the sum of unary and pairwise potential functions

$$\mathcal{E}(\mathbf{x}) = \sum_i \psi_i(x_i) + \sum_{(i,j)} \psi_{ij}(x_i, x_j) \quad . \tag{3.12}$$

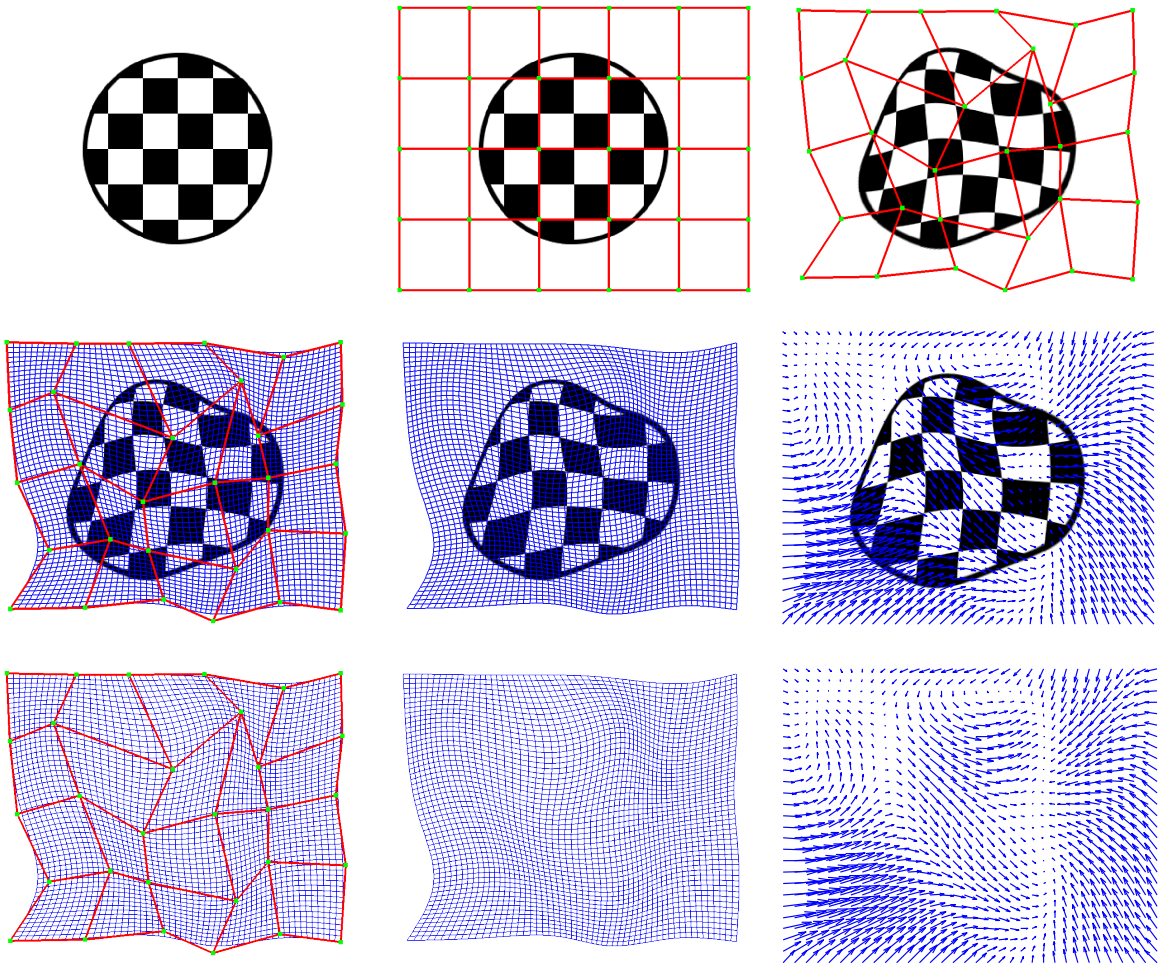


Figure 3.6: Free-form deformations. An example image (top-left) is embedded in a regular FFD grid (top-middle). By manipulating the control points (green), the shape of the ball-like object can be changed. The resulting deformations are smooth which is visible in the blue deformation and displacement field. We use backward warping for the image transformation; this is why the actual object deformation appears to be inverse to the deformation field.

We propose to encode the likelihood energy, i.e. the similarity measure, on the unary terms and the prior energy, i.e. the regularization, on the pairwise terms. This seems a rather natural choice and follows the vision examples from Chapter 1, but it is not the only possible choice as we will see later. Remember, the unary terms evaluate the label assignment x_i independently of all other assignments, while the pairwise terms consider two random variables, simultaneously.

3.4.3.1 Efficient Likelihood Approximation Scheme

So what is the likelihood energy for moving a control point? Remember, that a control point i has only local support on a sub-domain Ω_i centered at i. We define the unary

potentials based on a local similarity measure $\mathcal{S}_i$ as follows

$$\psi_i(x_i) = \mathcal{S}_i(I, J, \tilde{D}, x_i) \quad , \tag{3.13}$$

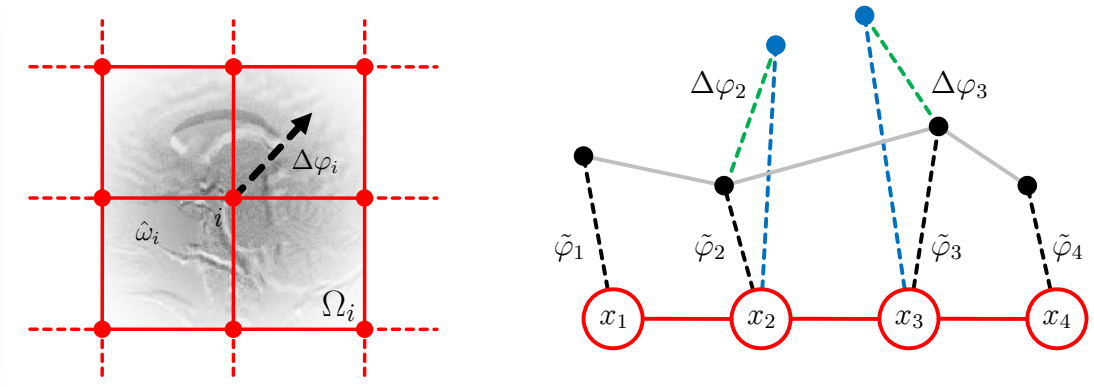
where $\tilde{D}$ is the current displacement field (from the previous iteration and initially set to zero). The similarity measure evaluates only within Ω_i how well the images I and J would be aligned if control point i is displaced by x_i. Note that I can correspond to a warped version of the original source image, obtained in previous iterations. As an example, let us have a look on how the local SAD measure looks like. We define it as

$$\mathcal{S}_i(I, J, \tilde{D}, x_i) = \sum_{p \in \Omega_i} \hat{\omega}_i(p) \, |I(p + \tilde{\delta}_p + x_i) - J(p)| \quad . \tag{3.14}$$

This definition needs some further explanation. The intensities of the source image I are read out after adding the displacement x_i, so we get $I(p + \tilde{\delta}_p + x_i)$. In fact, this is equivalent to a translation of the (warped) image patch Ω_i. But why do we add the update x_i equally on every image point? The point is, that the actual new displacement for an image point p is unknown, since it depends on more than one control point. But the unary term considers only the displacement of control point i. Indeed, from Equation (3.11) we could compute the resulting point displacements by moving control point i additionally with x_i and assuming that all other control points remain unchanged. But even then, the local similarity measure would only be an approximation. The problem is that we do not know whether the other control points will move or not. Whatever we assume, it will always be an approximation to the real energy. So, assuming a translation of the whole patch – which corresponds to the assumption that all control points take the same label – is just as well justified as any other assumption. But it has the important advantage that it is very fast to evaluate without the need for computing the displacements of image points.

In order to improve the approximation, we added weighting functions $\hat{\omega}_i$ which act similarly as the weighting functions of the control points; here $\hat{\omega}_i$ determines the contribution of image points p to the local similarity measure. Points closer to a control point should contribute more, since they are also more influenced by the control point displacement. We could choose the same weighting functions as the ones used for the transformation, e.g. cubic B-splines. In practice, it turns out that simple linear functions are usually sufficient. The concept of the weighting functions on the local domain is illustrated in Figure 3.7.

We should note that the before mentioned likelihood evaluation only works for pointwise similarity measures. Region-based measures such as the CC or MI extract further statistics from the intensity domain, and a weighted summation of pointwise distances does not work. In case of region-based measures, we follow a slightly different but still efficient approach. We employ local versions of the similarity measures. So, for instance the evaluation of the CC measure is simply restricted to the local domain Ω_i. This works extremely well. The same can be done for the MI based measures, but here we also have to determine local joint intensity distributions. In that case, it depends on the size of the local sub-domains how meaningful the local versions of MI are. With decreasing domain size, the statistics drawn from less intensity samples are degraded. In Chapter 4 we



(a) Likelihood (b) Prior

Figure 3.7: Illustration of the first-order energy terms. The local likelihood evaluation is illustrated in (a). At control point i only the local domain Ω_i is considered. Additionally, a weighting function $\hat{\omega}_i$ (white fade-out) controls the influence of image points for point-wise similarity measures. The prior term is shown in (b). For the fluid-like regularization (cf. Equation (3.16)) only the (green) updates $\Delta\varphi$ are considered. For the elastic-like regularization (cf. Equation (3.17)) also the previous (black) displacements $\tilde{\varphi}$ are considered which yields a penalty on the full (blue) displacements.

will present a solution to this problem which is based on the so-called pointwise mutual information (PMI) (cf. Section 4.1.2). For PMI our weighted evaluation can be applied.

One could ask, why we choose to encode the likelihood energy on the unary terms if it comes with the problem of approximation. First, the approximation is not too bad and we are able to achieve highly accurate registration results, as we will see later in Chapter 4. Second, our approach is extremely efficient from a computational perspective. In fact, we can compute the local similarity measures for all k control points and a certain displacement in just one loop over the image points. We simply apply the displacement under consideration as a global translation to the source image. Then we iterate once through the whole domain Ω and assign the local costs with the respective weightings to the k different unary terms. The weighting functions $\hat{\omega}_i$ take care that only points p within Ω_i are considered for control point i; outside the sub-domain the weighting is simply set to zero. This procedures allows us to very quickly compute an energy look-up table with $k \times |L|$ values, where $|L|$ is the number of labels. There are only $|L|$ global translations and loops over the source image. We call this procedure the *efficient likelihood approximation scheme* (ELAS).

3.4.3.2 Local Smoothness

The second part of the registration energy is the regularization term. We have previously mentioned the need for regularization in order to render the optimization problem well-posed. Adding a regularization term to the energy is one possibility to achieve that. However, actually the fact that we are using free-form deformations, and that the number

of control points is commonly much lower than the number of image points is also some kind of explicit regularization. The number of unknowns has been dramatically decreased compared to the number of observations, i.e. the image intensities. The problem is well-posed and additionally the FFDs guarantee smooth deformations. So what is the point of adding an implicit regularization in terms of an energy?

The motion of control points, and thus the registration, is driven by the likelihood, i.e. the similarity measure. Now, the problem is that in general not all areas in the images contain sufficiently enough structure to obtain a good alignment solely based on an intensity based similarity term. Homogeneous, noisy or corrupted areas might benefit from an additional regularization term. The idea is, in areas with prominent structures or a lot of texture the likelihood term will dominate and control the motion of the control points. This motion could then guide control points where the likelihood is less discriminative and reliable. To this end, we can encode local regularization terms $\mathcal{R}_{ij}$ on the pairwise potential functions as follows

$$\psi_{ij}(x_i, x_j) = \alpha_{ij}\, \mathcal{R}_{ij}(x_i, x_j) \quad . \tag{3.15}$$

The scalar α_{ij} is a weighting factor which controls the influence of the regularization term on the total energy. It can vary locally which is interesting if additional information is available. Otherwise we can set a fixed value for all pairwise terms. One possibility to achieve local smoothness is for instance by using the following definition

$$\mathcal{R}_{ij}(x_i, x_j) = \frac{\|x_i - x_j\|}{\|i - j\|} \quad . \tag{3.16}$$

This term penalizes variations between the displacements of neighboring control points i and j. It is a discrete approximation of a gradient penalty and favors smooth updates for the displacement field. However, only the updates are penalized. If we want to regularize the overall deformation, we need to consider the control point displacements from previous iterations, denoted by $\tilde{\Phi} = \{\tilde{\varphi}_1, ..., \tilde{\varphi}_k\}$. This can be done by defining the regularization term as

$$\mathcal{R}_{ij}(x_i, x_j) = \frac{\|(\tilde{\varphi}_i + x_i) - (\tilde{\varphi}_j + x_j)\|}{\|i - j\|} \quad . \tag{3.17}$$

Depending on whether Equation (3.16) or (3.17) is employed, the alignment of the images will behave differently. The penalty on the updates results in a fluid-like registration; every update is smooth, but the final deformation can take almost any form. In contrast, the penalty on the full deformation considering the previous displacements results in an elastic-like registration; the final deformation is much more constrained with respect to the initial one. The evaluation of both terms is illustrated in Figure 3.7(b).

The two presented regularization terms are probably the most basic ones. Others could be based on robust penalty functions by truncating or thresholding the energy at a certain level to allow discontinuities in the labeling – this is similar to the prior in the restoration problem. In our application chapter we will present some alternatives and compare the behavior of different regularization terms.

3.4.4 The Higher-Order CRF Model

Our second approach for non-linear registration is based on a higher-order conditional random field formulation. The term *high-order* reflects the fact, that we employ a model with cliques of size larger than two. To this end, we make use of triangulation meshes as a model for piecewise-affine transformations (cf. Figure 3.8). This approach has been recently presented in our work on optical flow estimation [38]. In contrast to the first-order model, the likelihood energy is modeled exactly without approximations. The formulations differs depending on the dimensionality of the images to be registered. In the following, we will focus on the formulation for the 2D case and briefly discuss the differences for 3D.

In many registration scenarios one can make the following two observations: (i) often the scene contains mainly solid objects, which might translate, rotate, and/or scale from one image to another, (ii) the motion of non-solid objects (such as textiles or tissues) can be represented by sufficiently many local affine motions. A transformation model that fits quite nicely to these observations are triangulation meshes. Assuming the image domain Ω is covered by a mesh of triangles. Each triangle is defined by three control points i, j, and k (the triangle vertices). Similar to FFDs, we obtain a dense displacement field D by displacing the control points. Following Equation (3.11), we can compute the displacements of image points p lying within a triangle (ijk) from a linear combination of control point displacements $\Phi_{ijk} = \{\varphi_i, \varphi_j, \varphi_k\}$ as

$$D(p) = \omega_i \varphi_i + \omega_j \varphi_j + \omega_k \varphi_k \quad . \tag{3.18}$$

Here, $(\omega_i, \omega_j, \omega_k)$ are the barycentric coordinates of point p with $\omega_i + \omega_j + \omega_k = 1$. In such a way, every triangle is representing a local affine warp T_{ijk} with six degrees of freedom (i.e. three 2D displacements) on the triangular sub-image I_{ijk}. Again the random variables $\mathbf{X}$ are associated with the updates on control point displacements $\Delta\Phi$ and labels correspond to displacements, i.e. $x_i \equiv \Delta\varphi_i$. The registration energy is encoded in the energy of a higher-order random field, which is generally defined as

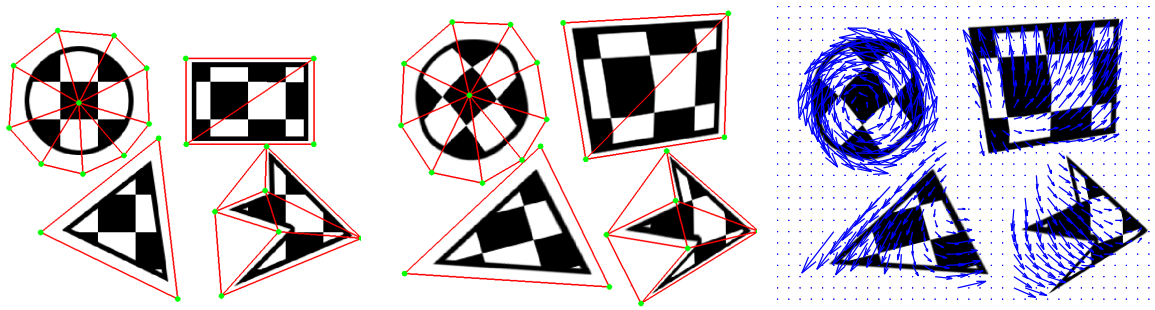
$$\mathcal{E}(\mathbf{x}) = \sum_{C \in \mathbf{C}} \psi_C(\mathbf{x}_C) \quad . \tag{3.19}$$

Here, $\mathbf{x}_C$ is the sub-labeling for the variables within clique C. Now, the task is to define the set of cliques and the posterior energy of the field.

3.4.4.1 Triangulation-Based Likelihoods

We define the likelihood energy on triple-clique potential functions. Each triangle (ijk) constitutes a clique and a similarity measure is evaluated for the warped triangle which occurs when labels x_i, x_j, and x_k are assigned. The similarity measure is considered between the warped sub-image I'_{ijk} and sub-image J_{ijk}. The corresponding potential function is simply defined as

$$\psi_{ijk}(x_i, x_j, x_k) = \mathcal{S}(I'_{ijk}, J_{ijk}) \quad . \tag{3.20}$$



(a) Triangulation (b) Deformation (c) Displacement field

Figure 3.8: Piecewise affine motion model based on triangulations. Each object is embedded in a triangulation mesh and can be separately transformed.

The great advantage of this formulation is that the likelihood energy is evaluated exactly without approximations. The displacements for all image points considered in the potential function are known since they are fully defined by the three labels respectively displacements of the three triangle vertices. The drawback is that we cannot employ a fast evaluation scheme as the one used for the first-order model. Indeed, the evaluation is exact but less efficient with respect to computation time.

If we want to extend this approach to 3D, we need to consider tetrahedralizations instead of triangulations. The likelihoods would then be based on quadruple-cliques defined for tetrahedral sub-volumes.

3.4.4.2 Geometric Regularization

Triangles covering homogeneous regions might lead to unreliable motion estimates. There are several ways for employing a regularization on the mesh. Note that our higher-order formulation implicitly has some kind of mesh regularization. Triple cliques of neighboring triangles will automatically have two control points in common, the ones on the common edge. Additionally, we propose a simple yet effective explicit regularization which is also based on triple-clique potential functions. We call it the *angle deviation penalty* (ADP). The ADP is defined as

$$\psi_{ijk}(x_i, x_j, x_k) = \alpha_{ijk} \|(\gamma_i, \gamma_j, \gamma_k) - (\gamma'_i, \gamma'_j, \gamma'_k)\| \quad . \tag{3.21}$$

The term penalizes the change between the initial angles $(\gamma_i, \gamma_j, \gamma_k)$ and the angles of the warped triangle $(\gamma'_i, \gamma'_j, \gamma'_k)$. The ADP is invariant to similarity transformations (i.e. all transformations containing only translation, rotation, and isotropic scaling). The weighting α_{ijk} controls the influence of the regularization energy when added to the likelihood energy. A different term, based on quadruple-cliques, which is penalizing non-affine motions between neighboring triangles, is considered in our optical flow paper [38] and therein also compared to the one presented here.

The extension to 3D is straightforward; the ADP for the four angles in a tetrahedron is simply encoded on quadruple-cliques.

3.4.4.3 Mesh Construction

A nice property of triangulation meshes is their great adaptivity and flexibility. The control points can be located anywhere in the image domain and thus the meshes can be adapted to the structures and objects visible in the images. A cleverly devised placement of triangles can yield more realistic transformation models in particular in areas with motion boundaries (cf. Figure 3.8). There are many ways how to obtain suitable, and in best case, automatically generated triangulations. In [52], we have investigated an approach which is based on control point placement at dominant points of intensities and compared this to free-form deformations with uniformly distributed control points. Recently, we have looked into more sophisticated approaches for mesh construction based on object and motion segmentation [38]. A very promising direction is the separation of motion layers and defining individual meshes for each layer; a direction which is followed in [38] and will be extended in future work.

3.4.5 Discrete Label Space and Refinement Strategies

Our two approaches for non-linear registration are both based on the estimation of labelings corresponding to control point displacements. So far we did not really discuss how this is done. Here, we make use of recent discrete optimization techniques, because they are able to obtain strong, sometimes globally optimal, solutions, at least within a discrete search space. Strong solutions are important, since they yield good updates on the transformation and result in efficient minimization of the original registration energy. The label set L defines our search and solution space in every iteration. The problem which arises here is the fact that our parameters, i.e. the displacements of the control points, live in a continuous space, so actually we would need the label space to be $L \subseteq \mathbb{R}^d$ (where d is the dimensionality of the images). But for discrete optimization the set of labels has to be discrete. So, how can we efficiently sample this space and form a set of discrete labels?

It is important to find a good compromise. On the one hand, a small number of labels allows fast optimization. On the other hand, a too sparse sampling of the continuous space may lead to inaccurate registration. To this end, we propose a *label space refinement strategy* which allows us to keep the number of labels small, but also to achieve high-accurate results. We commonly employ a sparse sampling with a fixed number of samples s. We uniformly sample displacements along certain directions up to a maximum displacement magnitude $\delta_{\max}$. The total number of labels is then $|L| = h \cdot s + 1$ including the zero-displacement and h is the number of sampling directions. Now, in every iteration of the registration process, we determine the optimal update by solving a discrete labeling on the current set of labels. If the update is successful, which means the energy decreases sufficiently, we keep the current set of labels for the next iteration. If the energy did not decrease, which means we cannot find any better alignment within the current search space, we refine the label space by rescaling every displacement in L by a factor $0 < f < 1$. The next iteration is then performed on this refined search space.

The number and orientation of the sampling directions depends on the dimensionality of the registration. The simplest possibility is to sample just along the main coordinate

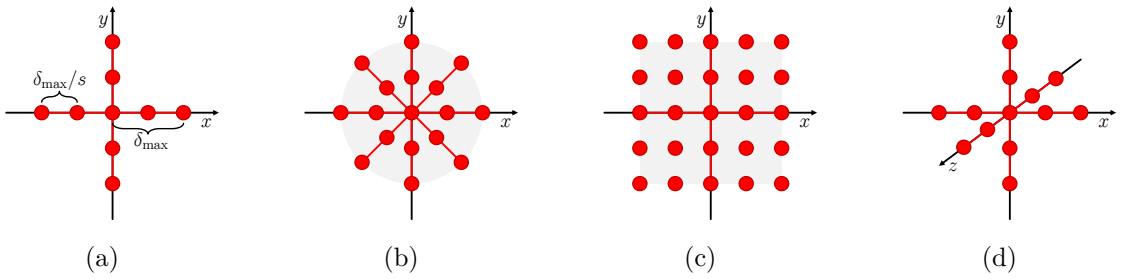


Figure 3.9: Discrete label spaces. The simplest strategy is to simple along the main coordinate axes as shown for 2D in (a) and 3D (d). We commonly employ sparse sampling such as shown in (b). A dense label space can be defined by uniform sampling as shown in (c).

axes, i.e. in positive and negative direction of the x-, y-, and z-axis (in case of 3D). Additionally, we can add samples for instance along diagonal axes. In 2D we commonly prefer a star-shape sampling, which turns out to be a good compromise between the number of samples and sampling density. In our experiments we found that also very sparse samplings, e.g. just along the main axes, gives very accurate registration results but comes with an increased number of iterations until convergence. The sampling strategies are illustrated in Figure 3.9.

3.4.5.1 Optimization

The remaining question in this part about non-linear registration is: how do we obtain the labelings? We need to employ one of the algorithms reviewed earlier in Chapter 2. Ok, but which one? What is the best optimization technique for our specific formulation? To answer this question, it is necessary to analyze some of the model properties. Both models, the first-order and the higher-order model, have one property in common: their energies are not necessarily *submodular* which has an impact on the applicability of some algorithms based on graph-cuts (cf. Section 2.4.1.1 in Chapter 2).

Regarding the higher-order model it should be quite clear to see that we cannot expect the clique potentials based on an intensity-based similarity measure to be submodular. The potential functions are simply too general.

For the first-order model it is not so obvious to see why the energies should not be submodular. The essential part here are the pairwise terms – remember, that unary terms are submodular by definition. If we want to use α-expansion with classical graph-cuts [18], the pairwise terms need to satisfy the following condition: $\psi_{ij}(\alpha,\alpha) + \psi_{ij}(\beta,\gamma) \leq \psi_{ij}(\alpha,\gamma) + \psi_{ij}(\beta,\alpha)$ for all labels $\alpha,\beta,\gamma \in L$. And here, the devil is in the details. Whether the pairwise terms fulfill this condition depends on the type of regularization, *fluid-like* or *elastic-like*, but also on the exact computation of the penalties. For instance, the fluid-like definition where only updates are penalized is submodular, if absolute vector differences are considered as in Equation (3.16). However, if we employ a quadratic version which is defined as

$$\mathcal{R}_{ij}(x_i,x_j) = \left(\frac{\|x_i - x_j\|}{\|i-j\|}\right)^2 \quad , \tag{3.22}$$

the pairwise terms are not submodular anymore. It is easy to construct an example to show this (e.g. with $\alpha=(2,0)^\top$, $\beta=(-2,0)^\top$, $\gamma=(3,0)^\top$).

The elastic-like definition is non-submodular, no matter if the absolute or quadratic version is used. It considers the displacements of previous iterations, and therefore it cannot be ensured that the condition holds. Consider a simple example: the current displacements of two neighboring control points are $\tilde{\varphi}_i=(1,0)^\top$, and $\tilde{\varphi}_j=(2,0)^\top$, their distance is assumed to be one, and the three labels correspond to updates $\alpha=(1,0)^\top$, $\beta=(2,0)^\top$, and $\gamma=(-1,0)^\top$. If we evaluate Equation (3.17) we get $\psi_{ij}(\alpha,\alpha)+\psi_{ij}(\beta,\gamma)=1+2$ and $\psi_{ij}(\alpha,\gamma)+\psi_{ij}(\beta,\alpha)=1+0$, so $3\not\leq 1$.

In conclusion, in most cases we will have to deal with functions which are non-submodular, so the optimization algorithm must be able to handle this class of energies. For the first-order model we employ the FastPD method [92, 93] which is known to be extremely efficient and it is able to handle our energies. Remember, similar to α-expansion, FastPD works by iterating over the set of labels (cf. Section 2.4.2.1 and Algorithm 1 in Chapter 2). FastPD is only applicable to first-order energies. So for our higher-order approach, we employ a different optimization technique. It is based on the QPBO algorithm [85, 128] in combination with higher-order clique reduction [65]. This combination allows to minimize non-submodular higher-order energies. Again, we perform the optimization in an α-expansion manner, where we iterate over the set of labels.

3.5 Linear Registration

Our models for non-linear registration exhibit certain analogies found in other random field models used in applications such as restoration or stereo (cf. Chapter 1). The random variables represent entities localized in the image domain and thus also the corresponding graph topology, and in particular the neighborhood system is related to the Euclidean distance of nodes. This is fundamentally different for the case of linear registration. The parameters of linear transformations act globally and cannot be assigned to certain points in the image domain. To this end, we propose a random field model which is very different from the previous ones. The first time we have presented this idea was in [46] and later more comprehensively in [163]. But before we come to the exact derivation of the model, let us discuss how we can represent linear transformations.

3.5.1 Parameterization

Linear transformations for d-dimensional images can be conveniently expressed in matrix form using homogeneous coordinates [53]:

$$A=\begin{bmatrix} \widehat{A} & \\ v^\top & 1 \end{bmatrix}\ , \tag{3.23}$$

where $\widehat{A} \in \mathbb{R}^{d\times(d+1)}$ and $v \in \mathbb{R}^d$. The new location of a (homogeneous) image point p is then defined via the matrix-vector-product as

$$T(p) = Ap \quad . \tag{3.24}$$

In the following, we focus on transformations up to *affine* (cf. Figure 3.4) by assuming that $v = \mathbf{0}$. However, the model can be easily extended to *projective* transformations. Affine transformations have 6 degrees of freedom in 2D, and 12 in 3D. One possible way for parameterization is simply to take every single entry of $\widehat{A}$ as one parameter. However, this not very intuitive since the majority of the entries does not have a direct geometrical interpretation. Instead, we employ a parameterization in which the affine transformation is decomposed into a set of matrices as

$$A = M_t R_\phi R_\theta^{-1} D_s R_\theta \quad . \tag{3.25}$$

Here, M_t represents a translation, R_ϕ a rotation, and $R_\theta^{-1} D_s R_\theta$ represents the shearing component. For the shearing, R_θ is a rotation and D_s is a diagonal matrix, representing anisotropic scaling. The single matrices in Equation (3.25) can be directly also represented by the respective parameters [53]. In 3D case, the rotation is parameterized by Euler angles. The resulting sets of parameters for the 2D and 3D case are

$$\Phi_{\text{2D}} = \{t_x, t_y, \phi, s_x, s_y, \theta\} \quad , \tag{3.26}$$

$$\Phi_{\text{3D}} = \{t_x, t_y, t_z, \phi_x, \phi_y, \phi_z, s_x, s_y, s_z, \theta_x, \theta_y, \theta_z\} \quad . \tag{3.27}$$

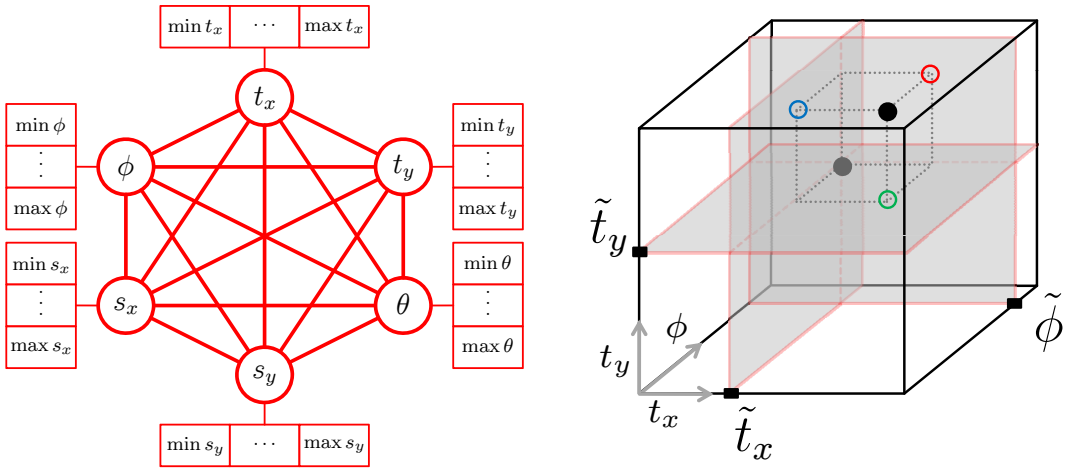
When we do not need to differentiate between the actual meaning of the parameters, we simply write $\Phi_{\text{2D}} = \{\varphi_1, ..., \varphi_6\}$ and $\Phi_{\text{3D}} = \{\varphi_1, ..., \varphi_{12}\}$ with $\varphi_i \in \mathbb{R}$.

3.5.2 The Highly-Connected First-Order CRF Model

Very similar to the non-linear case, the task of linear registration is now formulated as a parameter estimation problem where we need to find the optimal values for the parameters in Φ. As already mentioned, this can be achieved by iteratively seeking optimal updates $\Delta\Phi$ where each $\Delta\varphi_i$ is associated with a random variable X_i, and labels are equivalent to updates, i.e. $x_i \equiv \Delta\varphi_i$. Note, in contrast to the non-linear case, here the labels are one-dimensional which, at first glance, seems to make the problem a little bit easier. Additionally, in case of linear registration there is no need for regularization. Due to the relatively small number of parameters – compared to the number of parameters in non-linear registration – the energy formulation solely based on a similarity measure is well posed and we can write the minimization problem as

$$\hat{\Phi} = \arg\min_{\Phi} \mathcal{S}(I \circ T_\Phi, J) \quad . \tag{3.28}$$

So, the task is now to define a random field energy $\mathcal{E}$ which matches Equation (3.28); and it turns out that this is not straightforward. The main challenge lies in the dependency of the individual parameters. As already mentioned, each parameter has a global effect on the image transformation. For instance, optimizing the parameters independently is



(a) (b)

Figure 3.10: Highly-connected CRF model for linear registration. The graph topology for 2D affine registration is shown in (a). Each variable has its own set of labels. Exemplary visualization of the energy approximation via pairwise terms for the 2D rigid case is shown in (b). Here, t_x, t_y, ϕ denote the translation and rotation parameters, with initial values $\tilde{t}_x$, $\tilde{t}_y$, $\tilde{\phi}$. The evaluation of the original energy at the parameter point $\Phi = \tilde{\Phi} + \Delta\Phi$ (black) is approximated by the sum of the energy evaluations at the projections of Φ to the 2D subspaces (red, green, blue). The subspaces are orthogonal and all pass through the initial point $\tilde{\Phi}$ (gray).

doomed to failure. But how can we appropriately encode this dependence in the labeling energy?

In Chapter 1 we have seen that conditional dependence between variables is modeled via the cliques. In fact, since all n parameters of an affine transformation depend on each other, we would need to define a random field with exactly one clique of order $n-1$, i.e. the clique compassing all variables. For the 2D registration we need a fifth-order clique, for 3D an eleventh-order clique. The corresponding random field energy based on this single clique is simply

$$\mathcal{E}(\mathbf{x}) = \psi(\mathbf{x}) \quad \text{with} \quad \psi(\mathbf{x}) = \mathcal{S}(I \circ T_{\Phi+\Delta\Phi}, J) \ . \tag{3.29}$$

We can use the higher-order minimization scheme based on QPBO [85, 128] and clique reduction [65]. However, this is not very efficient. The reduction works by introducing auxiliary nodes. Transforming a fifth-order clique into pairwise form – which is necessary for the QPBO graph cut – yields in worst-case 49 additional nodes, while the eleventh-order clique yields up to 9217 additional nodes [65]. The question is whether we can model the global dependency differently?

We propose a highly-connected CRF model which is solely based on pairwise cliques which can be efficiently optimized. Our random field energy is defined as

$$\mathcal{E}(\mathbf{x}) = \sum_{(i,j)} \psi_{ij}(x_i, x_j) \ . \tag{3.30}$$

The energy is similar to the classical first-order energy but without unary potential functions. The pairwise potentials are defined as

$$\psi_{ij}(x_i, x_j) = \mathcal{S}(I \circ T_{\Phi+\Delta\Phi^{ij}}, J) \quad , \tag{3.31}$$

where $\Delta\Phi^{ij} = \{\Delta\varphi_1, ..., \Delta\varphi_n\}$ is the update on the transformation parameters with entries defined as

$$\Delta\varphi_k = \begin{cases} x_i & , k = i \\ x_j & , k = j \\ 0 & , k \neq i, j \end{cases} \quad . \tag{3.32}$$

So, the each pairwise term evaluates the similarity measure when exactly two parameters are changed simultaneously. Now the trick is the following: we connect every variable with all others, yielding a fully-connected pairwise graph (cf. Figure 3.10(a)). Thus, the full conditional dependence between all variables is approximated via a highly-connected first-order conditional random field which does not need any reduction yielding auxiliary nodes. The minimization can be performed by employing any efficient optimization technique such as FastPD [92] or standard QPBO.

The energy in Equation (3.30) is an approximation to the original registration energy. A visual interpretation of this is illustrated in Figure 3.10(b). In [163], we could show that iteratively minimizing the approximation is efficient and yields a very good solutions for the linear registration problem defined Equation (3.28).

A nice property of our approach is the flexibility with respect to the type of transformation. If we want to allow only rigid registration, i.e. translation and rotation only, we can simply deactivate the variables for scaling and shearing by disconnecting the corresponding nodes from the graph. Graph modifications become an intuitive way of changing the behavior of the registration.

3.5.3 Discretization and Optimization

In order to make use of discrete optimization, again we need to define the discrete set of labels. Note, this time the variables take one-dimensional values only but still they live in a continuous domain, which means the label sets should actually be $L_i \subseteq \mathbb{R}$. We employ a very similar discretization strategy as the one used for the non-linear registration. For each parameter, we create its own L_i with a specific value range $r_i = [\delta_{\min}, \delta_{\max}]$ which is uniformly sampled with a fixed number of steps. So, all label sets have the same size, but cover a parameter specific range of values. This is necessary due to the different meanings of the single parameters such as translation, rotation, or scaling.

In all experiments [46, 164, 163] in which we have tested our linear registration approach, we used the FastPD algorithm [92]. Within the iterative registration loop, we also use the same label set refinement strategy as introduced earlier. If an update decreases the energy sufficiently, we keep the current set of labels for the next iteration. If not, we refine the label spaces by rescaling every value in L_i by a factor $0 < f < 1$. The registration is then continued on this refined search space.

3.6 Related Work, Discussion, and Outlook

We would like to conclude this chapter by a discussion which provides some more insights into our proposed registration framework based on random fields. If we look into the literature, in particular for non-linear registration, we notice that most previous approaches are based on continuous optimization methods. Variational approaches [56, 107] based on the famous formulation by Horn and Schunck [59] are commonly solved either by standard steepest-descent or more advanced (preconditioned) versions of gradient-descent optimization. Non-linear registration using free-form deformations – first proposed by Rueckert et al. [133] – is also commonly solved by variants of gradient-descent [74]. Last but not least, registration via demons [146, 151] uses the gradient of the likelihood energy as driving force for computing the iterative updates. There are many other approaches which are more or less similar to one of the three before mentioned ones. All of them have in common, that they require the differentiation of the energy function. This can be straightforward for rather simply energies, for instance based on an intensity difference measures and a simple regularization term, but it can be also quite tedious for more complex energies.

3.6.1 Gradient-Free Optimization

All our registration methods, the linear method as well as the non-linear methods, are what we call *gradient-free*. In discrete optimization, the energy function is evaluated directly for a variety of possible labelings; neither analytical nor numerical differentiation is required. This is of particular interest if we want to test a novel, complex energy function which might be difficult or even impossible to differentiate. In our framework, it is straightforward to "plug-in" new energy terms.

Another great advantage of the discrete formulation is connected to the strategy of optimization. While continuous methods, in particular gradient-descent based approaches, might easily get stuck into local minima of non-convex functions (cf. Figure 2.1), the inherent "large neighborhood search" of discrete methods can overcome this limitation. In every iteration, several possible solutions are evaluated and the one corresponding to the (approximately) optimal energy is chosen. Depending on the definition and size of the search range, discrete methods allow to "jump out" of local minima. Indeed, the optimality is bounded by the discretization, but with intelligent refinement strategy the accuracy of continuous methods can be achieved.

3.6.2 Search Space Control

Additionally, the explicit definition of search space can be another advantage. In continuous methods it is rather difficult to control the search space. For instance, if we consider a steepest-descent algorithm, the only parameter the user can explicitly control is the step size in every iteration. However, the step size with respect to the magnitude of the gradient of the energy function might not have a very intuitive meaning. In contrast, in discrete methods the user has full-control on the definition of the label space. The range and the resolution of the search space can be explicitly set, but also the orientation

and direction of the search can be controlled. In fact, our framework allows to make use of any sort of prior knowledge that might be available for the search range. For instance, we make use of a geometric property of free-form deformations which is known to yield *diffeomorphic* deformations [98, 132] by simply limiting the maximum displacement magnitude to a specific value. In case of linear registration, the definition of the label spaces for the individual transformation parameters also allows us to restrict the space of possible transformations. But also the other way around, we can for instance set the initial label space for the rotation parameter to $L_\phi = \{-180°, ..., +180°\}$. This allows a quasi global search for the rotation and the registration becomes much less sensitive to the initial alignment.

Compared to previous applications, such as restoration, stereo, or segmentation, we find that our optimization strategy is establishing a completely novel perspective on the applicability of discrete optimization. Commonly, discrete labeling has been used as follows: a set of discrete labels was defined which completely covers the solution space. The estimated labeling was then considered directly as the final solution. This is very different compared to our setting where it is infeasible to define a single set of labels. Instead, we successively define sets which represent the possible updates for the current solution. The original energy is minimized iteratively, instead of estimating one single labeling. We believe that this perspective might be beneficial for many other optimization problems which so far have not been considered to be solved by powerful discrete methods.

Computation time, in particular for the first-order non-linear method, militates in favor of discrete approaches. The efficient likelihood approximation scheme in combination with the FastPD optimization method yields an algorithm which is considered to be one of the most efficient registration methods using free-form deformations. In the editorial preface of the journal where we first published this method it is stated: "The paper by Glocker et al. [44] introduces a novel and efficient approach do dense, non-rigid 3D image registration that reduces the required computation time from hours to minutes for large 3D voxel sets".

3.6.3 Other Discrete Approaches

It is worth to note that there are some works by other researchers which employed discrete optimization for registration. One of the first works is by Roy and Govindu [131] who use a multi-label optimization method which is based on a single *st*-cut [130, 62]. Their application is optical flow estimation which is closely related to non-linear registration. A random variable is introduced for every pixel in a 2D image. The pixel displacements are found by solving separately for the magnitude and the orientation. The approach is non-iterative, i.e. single labelings are computed where the label sets try to cover the whole solution space; hence, the resulting flow estimates are rather inaccurate.

Another non-iterative approach for the task of medical image registration has been proposed by Tang and Chung [144]. They also introduce one variable per image point, this time in 3D, and define a single set of labels which samples a dense 3D displacement space. The optimization is based on α-expansion; a similar approach has been previously proposed for the 2D case in [18]. The method is computationally very inefficient and a

single registration of moderately sized volumes takes more than 20 hours of computation time.

A different approach is proposed by Shekhovtsov et al. [136]. They present a decomposed model where a random variable is introduced per pixel and per dimension. The likelihood term is then encoded on pairwise potentials and TRW-S [83] is employed as the optimization algorithm. An extension of the decomposed model to 3D was later presented by Lee et al. [97]. The theoretical advantage of this approach is the reduced number of labels while a wide range of possible displacements can be covered at the same time. However, it turns out that this approach is less efficient and less accurate compared to our first-order non-linear FFD approach based on iterative refinement [137].

3.6.4 Further Ideas

An interesting development and future extension to our framework could be based on discrete-continuous optimization via fusion moves [101]. In cases, where the gradient of the energy function is available this information could be used for high-accurate refinement after initial discrete optimization. But instead of employing a standard gradient-descent, we could directly generate labeling proposals out of the gradient but with different scalings, i.e. different step sizes. The fusion move optimization is then performed on a set of gradient proposals, and the optimal one is selected individually for each random variable under consideration of a regularization energy. A rudimentary version of gradient proposals in fusion moves has been proposed in [64].

Another promising direction for future work is based on our higher-order CRF model. Here, investigations towards object-specific, or in the medical domain organ-specific, transformation models could help to dramatically improve non-linear registration. Since different tissues and materials have different deformation properties, these differences should be reflected in the model, in the transformation as well as the energy model. Deformation properties are sometimes already known or could be learned from annotated data. Our method with arbitrarily placement of control points seems suitable to encode such local properties.

In Chapter 1, we already mentioned that whatever kind of mathematical problem we want to solve on a computer, at some point discretization is unavoidable. The only question is where and when? In this thesis, we have proposed a novel perspective to this issue by introducing discrete random fields as an efficient alternative for image registration where the discretization is performed very early, i.e. in the phase of modeling. Despite the challenges which arose and which had to be tackled, we have shown the great potential lying in this approach. The following and last chapter of this thesis is dedicated to some specific applications in which our methods have been successfully applied.

CHAPTER

FOUR

APPLICATIONS

This final chapter is dedicated to real world applications and experiments in which we have used one of the registration methods presented in Chapter 3. We will present several results from the medical and non-medical domain and also show some recent extensions to our framework. We start by some general experiments which have been conducted to demonstrate the performance and flexibility of our approach. Afterwards, we present a variety of medical scenarios such as motion compensation, atlas-matching, and image stitching in which we successfully applied our registration algorithms. Finally, we show our recent results for optical flow estimation in non-medical images.

4.1 General Experiments

We would like to start by presenting some experiments which are not tailored to specific applications. This involves a comparison of our discrete approach to state-of-the-art registration methods based on continuous optimization, the investigation of a novel similarity measure, and an evaluation of different regularization terms. Last but not least, we present an extension of our non-linear framework which allows to incorporate prior knowledge from training data sets.

4.1.1 Discrete vs. Continuous

When we talk about solving the inherent continuous problem of image registration by means of discrete random fields, a natural question arising in this context is with respect to registration accuracy. We conducted many experiments in which we compared the registration accuracy of continuous methods with our discrete approaches. For example, a comprehensive experimental study for our linear registration method is presented in [46, 163]. For our non-linear first-order MRF approach we have performed several tests with known and unknown deformations in [43, 44]. For different similarity measures and a set of synthetically deformed images we find that our discrete approach is always competitive in accuracy and often much faster [44].

Gray Matter	Affine			Gradient-Descent			First-Order MRF		
Image	*DICE*	*Sens*	*Spec*	*DICE*	*Sens*	*Spec*	*DICE*	*Sens*	*Spec*
Brain 1	0.7022	0.7679	0.9633	0.8205	0.8547	0.9800	0.8567	0.8936	0.9831
Brain 2	0.7267	0.7236	0.9792	0.8142	0.8125	0.9857	0.8468	0.8489	0.9878
Brain 3	0.6687	0.6047	0.9816	0.8054	0.8059	0.9823	0.8332	0.8194	0.9867
Brain 4	0.7270	0.7924	0.9703	0.8154	0.8524	0.9818	0.8535	0.9065	0.9833
Brain 5	0.6977	0.7341	0.9686	0.8041	0.8449	0.9782	0.8355	0.8787	0.9809
Brain 6	0.7078	0.6328	0.9852	0.8116	0.7615	0.9891	0.8415	0.8112	0.9889
Brain 7	0.7062	0.6793	0.9779	0.8308	0.8303	0.9848	0.8591	0.8725	0.9857
Average	0.7052	0.7050	0.9752	0.8146	0.8232	0.9831	0.8466	0.8615	0.9852
White Matter									
Brain 1	0.6484	0.6214	0.9842	0.7686	0.7296	0.9910	0.8344	0.7909	0.9944
Brain 2	0.6269	0.6335	0.9863	0.7225	0.6794	0.9929	0.7962	0.8031	0.9924
Brain 3	0.6097	0.5622	0.9887	0.7312	0.7212	0.9899	0.7855	0.7937	0.9909
Brain 4	0.6860	0.6881	0.9866	0.7879	0.8034	0.9900	0.8428	0.8195	0.9947
Brain 5	0.6372	0.6080	0.9853	0.7598	0.7231	0.9912	0.8329	0.8297	0.9921
Brain 6	0.6521	0.6477	0.9882	0.7338	0.8976	0.9808	0.7794	0.8659	0.9876
Brain 7	0.6430	0.5924	0.9884	0.7840	0.8102	0.9881	0.8262	0.8312	0.9916
Average	0.6433	0.6219	0.9868	0.7554	0.7664	0.9891	0.8139	0.8191	0.9920
Running Time	4 minutes			3 hours 50 minutes			8 minutes		

Table 4.1: Deformable inter-subject brain registration. Given is the DICE score, the sensitivity, and the specificity for the alignment of the segmented gray and white matter of the brain. From left to right we give the results for the affine pre-alignment, the deformable registration using standard gradient-descent, and the results for our first-order MRF approach. Our approach is remarkably faster while yielding in average a higher accuracy for the alignment of the segmented structures. The size of each data set is $256\times256\times128$ voxels.

Here, we would like to present one particular experiment, namely the deformable inter-subject registration of MRI brain images. In order to detect differences in brain anatomy we registered 8 data sets in which the gray and white matter of the brain have been manually segmented. After registration it is possible to identify areas with large deformations and mark them as significant variations. The transformations estimated from the registration of the raw intensity data can also be applied to the segmentations. By determining how well the segmentations are aligned we can assess the quality of the registration. We selected one of the 8 data sets to act as the target image and the remaining 7 data sets have been registered to this target. The results are summarized in Table 4.1 and illustrated in Figure 4.1. More details can be found in [44]. The data is part of the Internet Brain Segmentation Repository (IBSR) provided by the Center for Morphometric Analysis at Massachusetts General Hospital[1]. We observe that our discrete approach is much more efficient in terms of computation time and we are able to obtain a similar (or even higher) accuracy for the alignment of the segmented structures compared to employing a standard gradient descent approach as described in [133].

[1] `http://www.cma.mgh.harvard.edu/ibsr/`

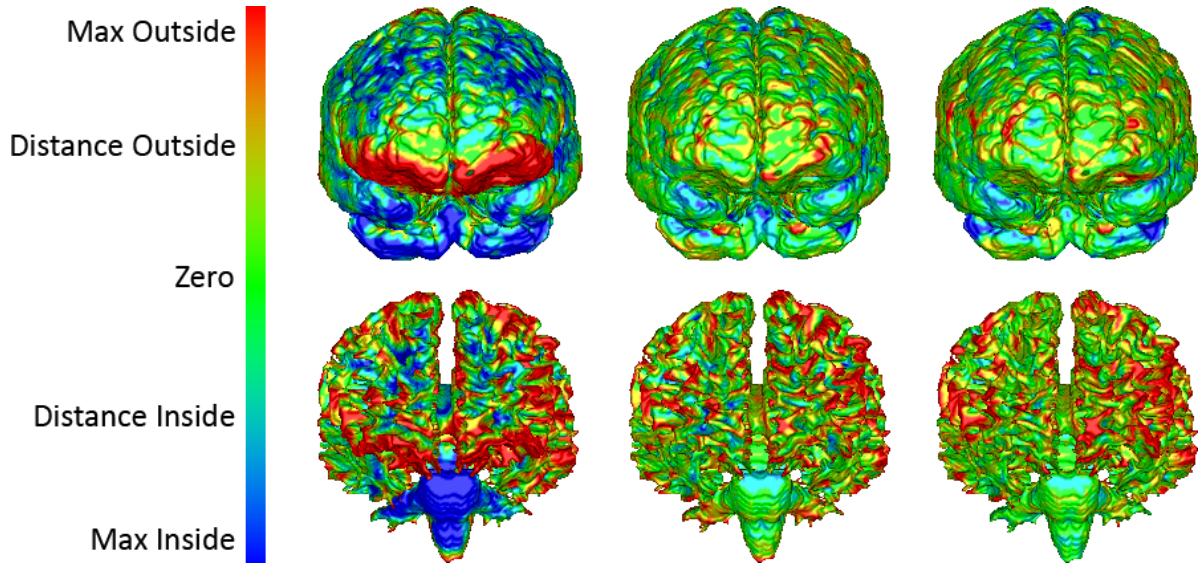


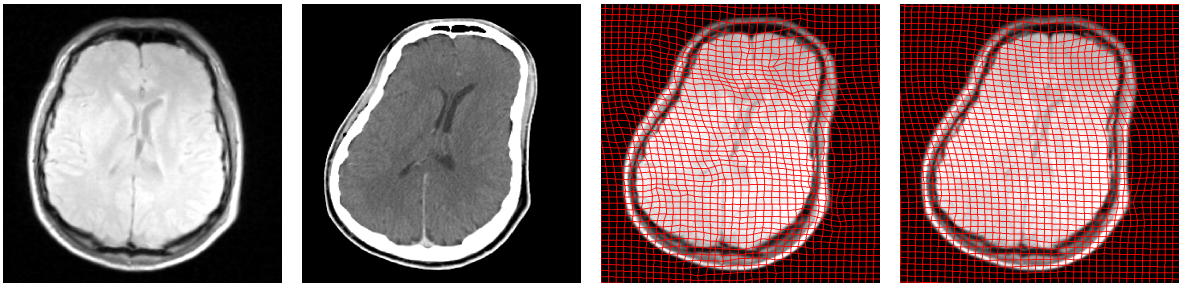
Figure 4.1: Deformable inter-subject brain registration. Color encoded visualization of the surface distance between the warped and expert segmentation after affine, gradient-descent, and our registration (from left to right) for the Brain 1 data set. The color range is scaled to a maximum and minimum distance of 3 mm. In some regions, the result for the gradient-descent approach seems to be slightly better. However, the actual average surface distance after registration for gray matter is 1.66, 1.14, and 1.00 millimeters and for white matter 1.92, 1.31, and 1.06 millimeters.

4.1.2 Pointwise Mutual Information

In the introduction of our non-linear first-order MRF approach (cf. Section 3.4.3) we already mentioned a problem which might arise in case of multi-modal registration. The problem is in the evaluation of the local statistical similarity measure. The local image statistics are calculated from local patches whose size decreases with an increasing resolution of the free-form deformation control grid. If the local domains are too small the statistics might become unreliable. This becomes an issue in particular for measures such as mutual information where joint intensity distributions are determined from the local domains. We tested an alternative approach for multi-modal registration based on so called *pointwise mutual information* (PMI) [123]. In every iteration, we compute only one global joint image histogram based on the current deformation. The mutual information is then calculated pointwise with respect to the pre-computed global intensity distributions $\rho_I(i)$, $\rho_J(j)$, and $\rho_{IJ}(i,j)$ as follows

$$\mathcal{S}_i(I, J, \tilde{D}, x_i) = -\sum_{p \in \Omega_i} \hat{\omega}_i(p) \log\left(\frac{\rho_{IJ}(I'_p, J_p)}{\rho_I(I'_p)\,\rho_J(J_p)}\right) \quad , \tag{4.1}$$

where $I'_p = (p + \tilde{\delta}_p + x_i)$ and $J_p = J(p)$. The great advantage of the pointwise calculation is that its statistical expression is less dependent on the actual resolution of the control grid. Similar to other pointwise measures, we can also employ weighting functions $\hat{\omega}$. Additionally, the computation of pointwise MI is much faster than its region-based counterpart since only one joint histogram has to be determined in case of PMI. In Figure 4.2 we show a comparison between PMI and region-based MI in a simple experiment. A 2D-MR image



(a) Source (b) Target (c) Region-based MI (d) Pointwise MI

Figure 4.2: Comparison of local region-based mutual information and pointwise mutual information. For very fine control grid resolutions the local statistics for region-based MI are less meaningful and yield unreliable control point displacements. In contrast, the resulting grid deformation in case of PMI is much smoother.

is registered to a 2D-CT image where the latter one has been synthetically deformed. We perform a hierarchical registration with increasing control grid resolution. We start with a 32 pixels control point spacing, then we reduce it to 16, and finally to 8 pixels spacing. In the finest resolution we can clearly see that for region-based MI the displacement estimates are becoming less reliable which is visible in the non-smooth deformation of the grid. In contrast, the resulting deformation when using PMI as the similarity measure is much smoother. Note, this is not an issue of regularization. In both cases, the weighting of the regularization was optimized. An increased regularization weight in case of region-based MI yields a much worse alignment of corresponding structures. The registration with PMI took about 8 seconds, while the region-based MI takes more than 19 seconds. We believe it is worth to follow-up these observations in future investigations.

4.1.3 Different Regularization Terms

Our first-order MRF approach for non-linear registration is extremely efficient from a computational perspective. However, we also know that first-order random fields in general have only limited capabilities for modeling more complex energy terms since they are restricted to pairwise interactions. This can become an issue if we want to employ regularization terms which go beyond penalizing first-order derivatives of the displacement field. Remember, we have introduced the absolute vector difference (cf. Equation (3.17) and its quadratic counterpart as approximations for first-order derivative penalties. The corresponding discrete filter mask is [–1 1]. They are encoded on the pairwise terms as

$$\mathcal{R}_{ij}(x_i, x_j) = \frac{\|(\tilde{\varphi}_i + x_i) - (\tilde{\varphi}_j + x_j)\|}{\|i - j\|} \quad , \tag{4.2}$$

$$\mathcal{R}_{ij}(x_i, x_j) = \left(\frac{\|(\tilde{\varphi}_i + x_i) - (\tilde{\varphi}_j + x_j)\|}{\|i - j\|} \right)^2 \quad . \tag{4.3}$$

The problem with these terms is that they also penalize linear transformations such as rotation or scaling. Indeed, we commonly perform a linear pre-alignment before running

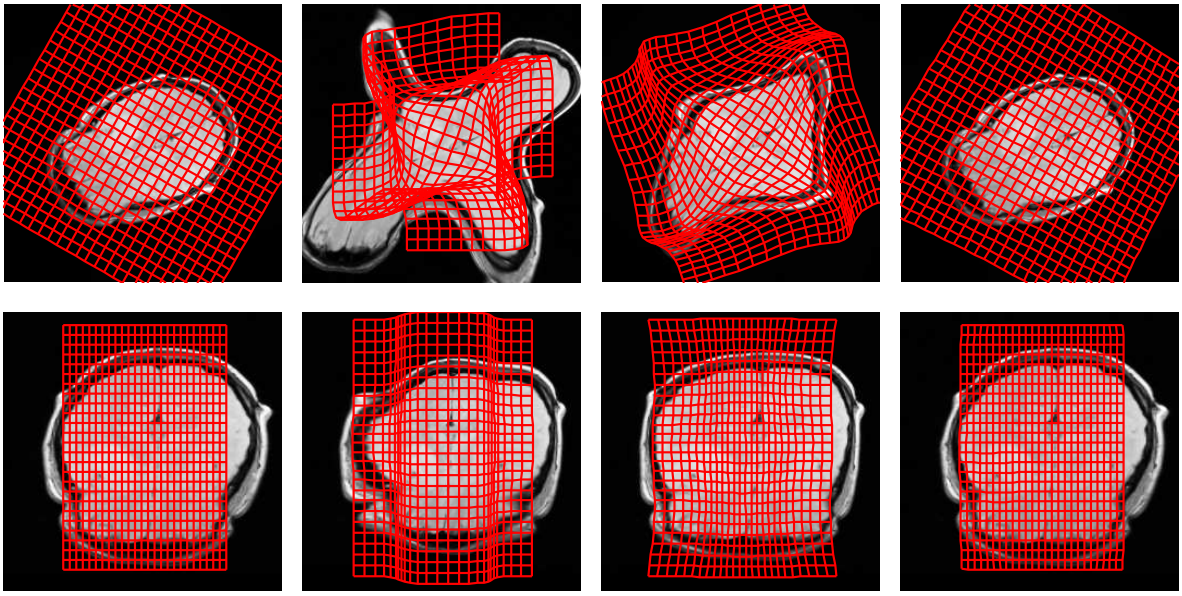


Figure 4.3: Comparison of different regularization terms for landmark-based registration. From left to right: ground truth transformation, and the results for absolute vector difference, quadratic vector difference, and approximated curvature penalty.

the non-linear registration, so most of the linear part of the transformation is already recovered and mainly non-linear deformations are left. However, in some applications it might be preferable to have a penalty term which is invariant to linear transformations. A penalty which has this property is based on second-order derivatives of the displacement field; but second-order derivatives require second-order interaction terms in form of triple cliques and cannot be encoded in a first-order random field energy. Hence, Kwon et al. [94] have proposed to use a (less efficient) second-order random field for registration; they encode the regularization on the triplet potentials following the discrete filter mask for second-order derivatives $[1 \;\; -2 \;\; 1]$ as follows

$$\psi_{ijk}(x_i, x_j, x_k) = \alpha_{ijk}\, \mathcal{R}_{ijk}(x_i, x_j, x_k) \quad , \tag{4.4}$$

$$\mathcal{R}_{ijk}(x_i, x_j, x_k) = \frac{1}{\delta^2} \left(\| (\tilde{\varphi}_i + x_i) - 2\,(\tilde{\varphi}_j + x_j) + (\tilde{\varphi}_k + x_k) \| \right)^2 \quad . \tag{4.5}$$

Here, α_{ijk} is the typical weighting parameter and δ is the distance between the control points. Due to the triplet potentials this approach is less efficient compared to our first-order model (cf. Chapter 2). In [41] we wanted to find out whether it is possible to define a regularization term which has similar properties as the above second-order penalty based on triple cliques while preserving the efficiency of a pairwise model. To this end, we have proposed what we call the *approximated curvature penalty* (ACP). The ACP is encoded on the pairwise potentials as follows

$$\begin{aligned} \mathcal{R}^H_{ij}(x_i, x_j) = \frac{1}{2\delta^2} & \left(\| \tilde{\varphi}_{i-1} - 2\,(\tilde{\varphi}_i + x_i) + (\tilde{\varphi}_j + x_j) \| \right)^2 \\ & + \left(\| (\tilde{\varphi}_i + x_i) - 2\,(\tilde{\varphi}_j + x_j) + \tilde{\varphi}_{j+1} \| \right)^2 \quad . \end{aligned} \tag{4.6}$$

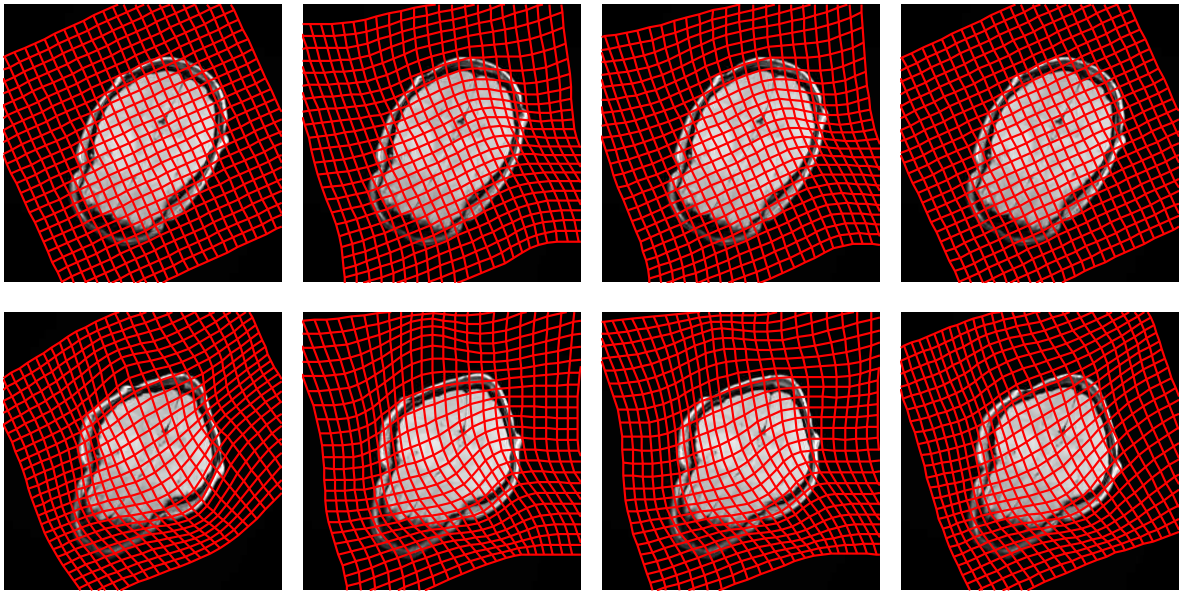


Figure 4.4: Comparison of different regularization terms for intensity-based registration. From left to right: ground truth transformation, and the results for absolute vector difference, quadratic vector difference, and approximated curvature penalty.

The idea of ACP is that we consider the displacements $\tilde{\varphi}_{i-1}$ and $\tilde{\varphi}_{j+1}$ of the two adjacent nodes of i and j without knowing there actual updates (cf. Figure 3.7(b)); hence it is only an approximation of the second-order derivatives. The above definition applies if i and j are horizontal neighbors. For the vertical case and for neighbors along the z-direction the definition is straightforward since only the indices of the two adjacent nodes have to be changed. The advantage of ACP compared to the second-order version in [94] is the efficient optimization. But how does it perform in practice and does it really have similar properties?

We have performed several experiments in order to investigate the behavior of ACP and also compared to the absolute and quadratic vector differences. In the first set of experiments we employed a likelihood function based on the Euclidean distance of K landmarks (instead of an intensity-based similarity measure). The landmarks correspond to the ground truth (linear) transformation between the two images. The likelihood term is

$$\psi_i(x_i) = \sum_{k=1}^{K} \hat{\omega}_i(p_k) \, \|(p_k + \tilde{\delta}_{p_k} + x_i) - q_k\| \quad . \tag{4.7}$$

The unary terms are the driving force for the control points close to the landmarks; minimizing the Euclidean distance will make sure that the corresponding landmarks will be aligned, perfectly. The question is what will happen to the rest of the displacement field which is solely following the forces of regularizations? We show the results for two different transformations, a rotation and a scaling, in Figure 4.3. We can clearly see how well the ACP behaves in case of linear transformations, while both other terms fail to recover the correct transformations. Another set of experiments was then performed

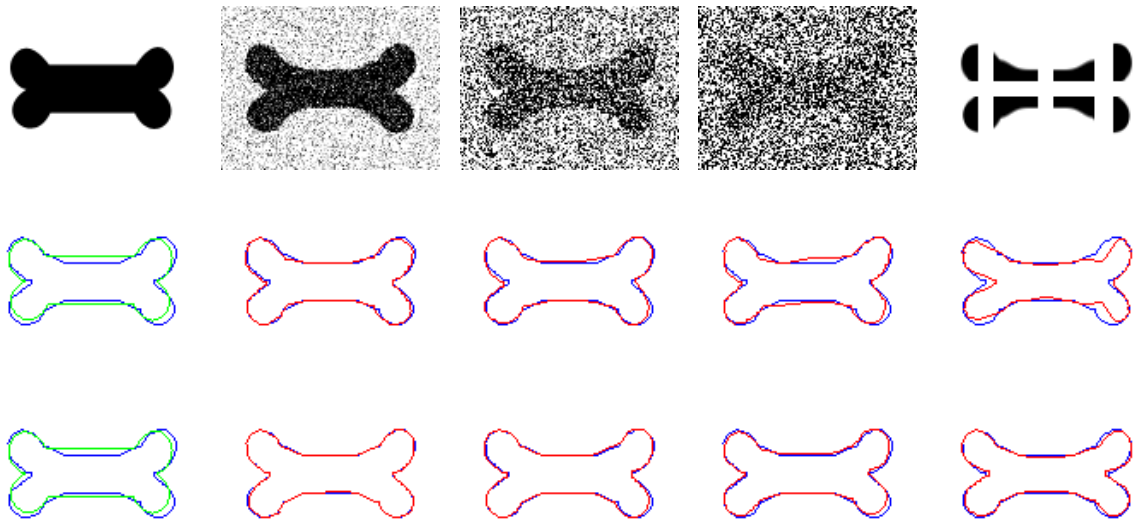


Figure 4.5: Learned deformation priors. The top row shows the source image (most left) followed by different target images with an increasing amount of noise or corruption. On the left in row two and three the initial alignment of the two object boundaries is shown (source in green and target in blue). The second row shows the results for an intensity-based similarity measure combined with conventional regularization. The boundary alignment is getting increasingly worse from left to right (the warped source boundary is shown in red). In contrast – even for the tough cases – the same similarity measure combined with a learned deformation prior can be used to properly align the two shapes (last row).

intensity-based, where the likelihood term was set to the SAD similarity measure. Again, we can see in Figure 4.4 that when a considerable fraction of the transformation is linear the ACP regularization term yields more reasonable displacement fields.

4.1.4 Learned Deformation Priors

Quite related to the above issue on regularization is our approach for incorporating learned deformation priors [39]. Here, we consider applications in which either repetitive motion patterns occur such as in cardiac imaging or registration scenarios in which similar deformations can be expected even if different images are considered. In these cases, the set of feasible deformations can be often represented by a compactly parameterized probability distribution which can be learned from a sequence of training data. Now, the idea is to encode this learned distribution in the prior energy of the random field such that deformations which have been seen before – the ones having high probability with respect to the learned distribution – are favored. This is of particular interest if the likelihood energy alone is not sufficient to drive the registration towards a good solution, which is often the case if the images to be registered exhibit a low signal to noise ratio or if the images are corrupted in certain areas.

Our approach presented in [39] works as follows: assuming we are given a set of N training examples (each example consists of a pair of images). Then we can perform N registrations each yielding one dense displacement field. The displacement fields can be parameterized via a set of FFD control points. For each control point we have a set of

N displacements – the ones from the training registrations. We estimate the underlying distribution over the set of displacements (e.g. via Gaussian mixture models [10]). Let us assume the FFD control grid has M control points, then we get M displacement distributions in total. Based on these distributions, we perform a control point clustering such that control points with similar distributions fall into the same cluster. Now, the interesting part is that instead of using a regular grid-like random field topology with a 4-connected neighborhood (in case of 2D), we define the topology based on the clustering. This means, that only control points within the same cluster are connected and thus conditionally dependent. The effect of this procedure is that regularization is only performed between control points which have similar distributions (obtained from the training). Particularly interesting is also the fact that learning-based regularization and the resulting non-regular random field topology can yield preservation of discontinuities between control points which are direct neighbors but with different motions. Since control points from different clusters are not connected, there will be no regularization term which otherwise might tend to oversmoothing. The details of our approach can be found in [39]. Here, we only show some visual results in Figure 4.5 for an experiment with synthetic deformations. We could imaging that learned deformation priors might be beneficial for instance in medical applications using ultra sound imaging where the quality of the images often prohibits the use of currently available intensity-based registration methods.

4.2 Medical Image Registration

In the following we present some medical applications in which our registration methods have been used. The first application is a linear registration problem, the rigid alignment of brain images from different modalities. As already mentioned earlier in such applications the main interest is to enrich the information about certain anatomical structures by fusing the data form different sources. The second application deals with an important component in computer-aided diagnosis of osteoarthrosis (OA), namely the segmentation of cartilage tissue in MR images. The loss of cartilage tissue is an indicator for the stage of OA. We demonstrate how we can make use of our non-linear registration, first to generate an atlas image of cartilage tissue, and second to automatically match the atlas to a new subject in order to obtain a segmentation. Our last application is whole body MR imaging in which non-linear registration is needed for simultaneous distortion correction and stitching of single MR images into one high-resolution scan. Other applications which are not covered here but where our methods have been used are registration of thoracic CT images [42], construction of statistical shape models [162], and simultaneous landmark and intensity-based registration for non-rigid brain registration [140].

4.2.1 Multi-Modal Brain Registration

In this experiment we demonstrate the performance of our highly-connected first-order CRF model for linear registration (cf. Section 3.5.2) in a real multi-modal rigid registra-

CT-MRI	Mean Error			Median Error			Maximal Error		
Protocol	*Simplex*	*Elastix*	*CRF*	*Simplex*	*Elastix*	*CRF*	*Simplex*	*Elastix*	*CRF*
MR-PD	2.067	2.226	2.078	2.005	2.018	1.986	4.052	6.310	3.884
MR-T1	1.275	1.334	1.286	1.259	1.230	1.154	2.873	3.172	3.003
MR-T2	2.053	2.085	1.856	1.979	1.950	1.853	4.271	4.192	3.647
Overall	1.789	1.870	1.729	1.743	1.815	1.739	4.271	6.310	3.884

Table 4.2: Quantitative registration results of the CT-MRI registration for three different methods including our highly-connected first-order CRF model. All errors are given in millimeters.

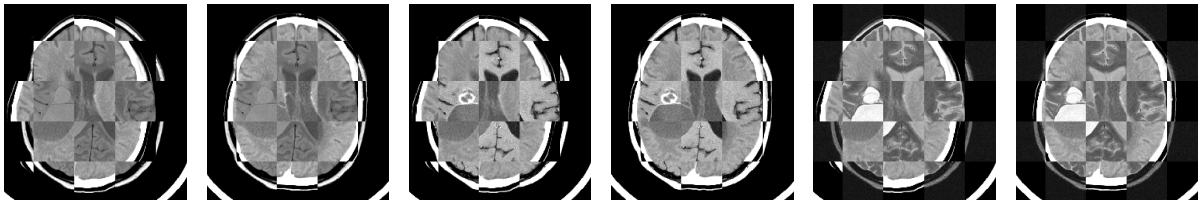


Figure 4.6: Checkerboard visualization of CT-MRI alignment before and after registration using our highly-connected first-order CRF approach. From left to right, the initialization and the registration results for MR-PD, MR-T1, and MR-T2 protocols.

tion scenario. The images are part of the RIRE project[2]. The task is the registration of CT brain images to MR images with different protocols, namely MR-PD, MR-T1, and MR-T2. As the similarity measure, we use the entropy correlation coefficient (ECC). We perform an extensive test on several patient datasets, for which the evaluation is performed remotely by the RIRE system.

In order to assess the accuracy of the proposed method, we perform the same tests also by two other methods. The first one is based on Downhill-Simplex optimization [118], the second one is the module of rigid registration from the *Elastix* toolkit [73] based on adaptive stochastic gradient descent [72].

We used all data sets from the RIRE data base, for which the MR-PD, MR-T1, and MR-T2 data sets are available, resulting in 10 patients with 3 registrations per patient. The tests show consistent performance of our discrete approach. Table 4.2 summarizes the results and shows that the proposed method slightly outperforms the other tested methods in terms of accuracy. Some visual results are shown in Figure 4.6. More details and further experiments can be found in our articles [46, 163, 164]. Therein, we also used our linear registration for the challenging task of 2D-3D registration, where a 2D projection image is registered to a 3D volume.

4.2.2 Atlas-Based Cartilage Segmentation

Our work on atlas-based segmentation [40] demonstrates one of the advantages of our discrete formulation, namely the great flexibility in adding new similarity measures. The medical motivation in this work is to obtain automatic segmentation of the cartilage tissue

[2] `http://www.insight-journal.org/rire/`

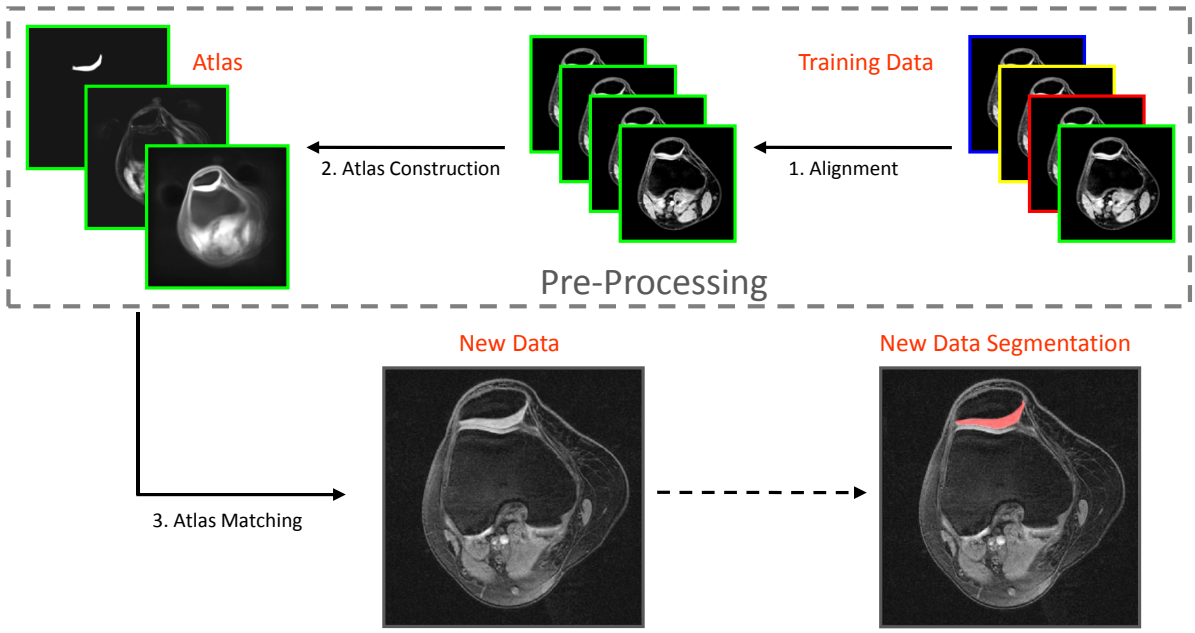


Figure 4.7: Atlas-based cartilage segmentation. Top row shows the atlas construction process. New data can be automatically segmented via atlas matching by non-linear registration with a specific matching criterion.

in MRI knee data. To this end, we make use of our first-order MRF method for non-linear registration with a specific matching criterion. The idea is the following: given a set of training images $I_1, ..., I_n$ in which cartilage tissue has been manually segmented (by an expert). From the training data, we construct what we call an *atlas* consisting of an statistical appearance image and average segmentation (cf. top row in Figure 4.7). For the atlas construction the training images need to be registered. Here, we follow the unbiased construction scheme proposed by Joshi et al. [67]: iteratively all images are registered to their average intensity image I_μ; after each iteration the average is recomputed. Just a few of such iterations are needed until the average image does not change anymore. Additionally to this average we compute a variance image I_{σ^2} containing the intensity variance at every image point. The estimated transformations from the registration of the training images are also applied to their segmentations allowing us to compute an average segmentation. All these steps are done off-line in a pre-processing step which has to be done only once.

Now, in the clinic when we acquire an MRI scan of a patient, we use the atlas data and register it to this new data set J. This allows us to obtain a fully-automatic segmentation of the cartilage by warping the average segmentation with the estimated transformation onto the patient's MR image. This is what we call *atlas matching* (cf. bottom row in Figure 4.7). Again, we use our general method for non-linear registration for this task but with an atlas-specific similarity measure which is defined as

$$\mathcal{S}_{\text{atlas}}(I_\mu, I_{\sigma^2}, J) = \frac{1}{|\Omega|} \sum_{p \in \Omega} \frac{(I_\mu(p) - J(p))^2}{\sqrt{I_{\sigma^2}(p)}} \quad , \tag{4.8}$$

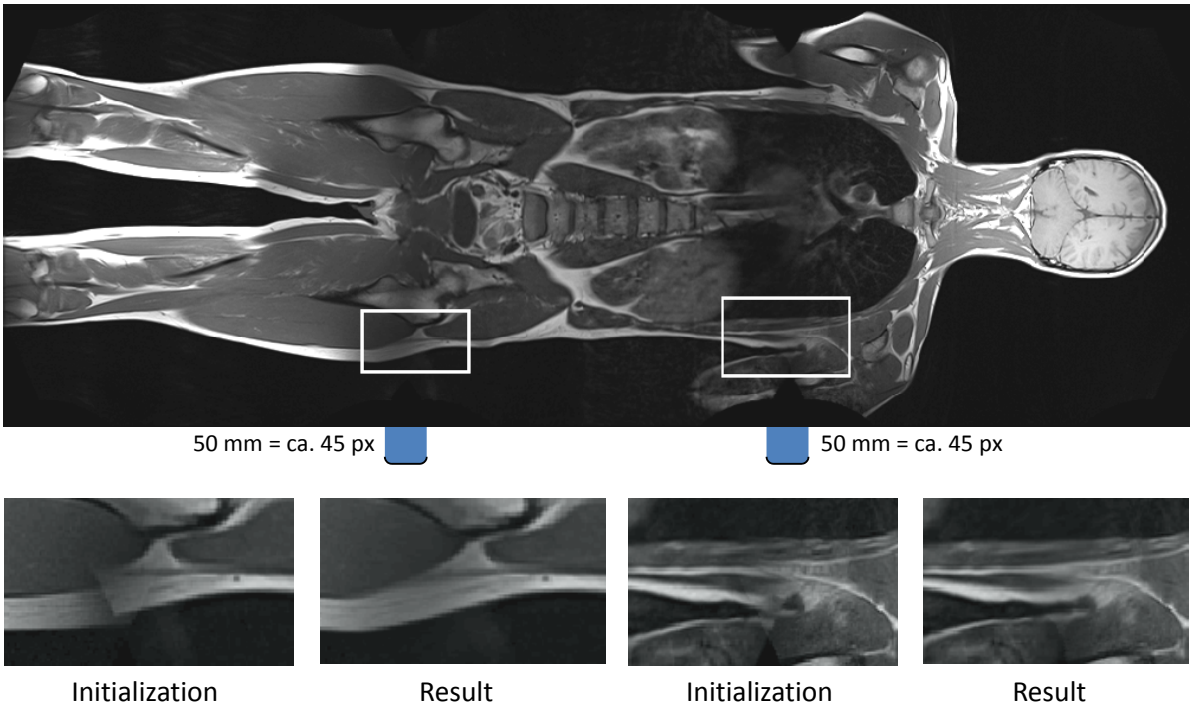


Figure 4.8: Image stitching for whole body MRI. Distortion and breaks in the overlap areas are corrected.

and the corresponding local similarity measure for our first-order MRF approach is then defined as

$$\mathcal{S}_i(I_\mu, I_{\sigma^2}, J, \tilde{D}, x_i) = \sum_{p \in \Omega_i} \hat{\omega}_i(p) \frac{\left(I_\mu(p + \tilde{\delta}_p + x_i) - J(p)\right)^2}{\sqrt{I_{\sigma^2}(p + \tilde{\delta}_p + x_i)}} \quad . \tag{4.9}$$

This measure simply assumes a Gaussian for the intensity distribution at every image point in the MR images. In our evaluation presented [40], we obtained quite good segmentation results with this approach. However, in cases where a simple Gaussian distribution cannot sufficiently represent the underlying data one might need to go towards more powerful statistical representations (e.g. Gaussian mixture models). But even then, it is straightforward to encode the corresponding similarity measure within our framework.

4.2.3 Image Stitching for Whole Body MRI

Our last application from the medical domain is the creation of high-resolution whole body MR images via image stitching and distortion correction [153]. Whole body imaging is an emerging application gaining enormous clinical interest, e.g. for screening. Commonly, the imaging is performed by acquiring several smaller high-resolution sub-images with sufficiently large overlap. After acquisition, the images are simply combined into a larger image and a blending operation (such as averaging) is applied in the overlap areas. The offset between the single images is known from the MRI device. The main drawback of such an approach is the long acquisition time – and time is one of the most expensive

resources in clinical environments – since many small sub-images need to be acquired in order to avoid distortion artifacts. The reason for this lies in the physical properties of MR scanners. If the field-of-view is increased towards the maximum of let's say 50cm the images might get distorted at the boundaries due inhomogeneities in the magnetic field. These are non-static distortions which cannot be corrected in advance and even depend on the subject being imaged [153]. However, a larger field-of-view would reduce the number of sub-images and thus decrease the overall acquisition time. Here, our non-linear registration comes into the game.

Given two high-resolution, large field-of-view sub-images I_1 and I_2 with an overlap area $\Omega_O = \Omega_1 \bigcap \Omega_2$ (known from the MRI device). In order to combine them into a larger image, we need to align these two images non-linearly due to the distortions in the overlap area. In contrast to regular registration scenarios, here we do not have designated target image. Both images are distorted, and none should remain fixed. To this end, we propose a strategy borrowed from the atlas construction scheme. We are seeking two transformations T_1 and T_2 simultaneously by registering the two images to their average. However, a classical average image is not appropriate due to the increasing amount of image distortion towards the boundaries. The average image is simply to blurry and almost meaningless in these regions. Therefore, we define what we call a *linear weighted average* where the idea is to account for the physical property of increasing distortion. Assuming that the boundary information of each image is less reliable, we would like to reduce its influence to the registration.

The linear weighted average image A is computed from the images I_1, I_2 as

$$A(p) = \begin{cases} f(p) & , p \in \Omega_O \\ I_1(p) & , p \in \Omega_1 \setminus \Omega_2 \\ I_2(p) & , p \in \Omega_2 \setminus \Omega_1 \end{cases} , \tag{4.10}$$

where f is a function computing the weighted average intensity in the overlap domain, defined as

$$f(p) = (1 - h(p)) \cdot I_1(p) + h(p) \cdot I_2(p) \ . \tag{4.11}$$

The linear function h takes increasing values between $[0, 1]$ along the stitching direction. Based on the linear weighted average we can define a similarity measure for the simultaneous registration. For instance the SAD criterion would be defined as

$$\mathcal{S}_{\text{stitch}}(I_1, I_2, A) = \frac{1}{2|\Omega_O|} \sum_{i=1}^{2} \sum_{p \in \Omega_O} |A(p) - I_i(p)| \ . \tag{4.12}$$

Similar to the atlas construction, we perform several iterations where in each iteration the linear weighted average is recomputed based on the current estimate for the two transformations. Encoding this registration scheme in our first-order MRF model is straightforward. The random field now consists of twice as much variables as in regular registration, since we have two free-form deformation control grids, one per image. Some visual results for a whole body scan originally consisting of three sub-images is shown in Figure 4.8. The two stitches have been computed separately. Our approach yields both much sharper transitions in cases of larger overlaps and correction of breaks and heavy distortions in case of very small overlap areas as shown in further experiments presented in [153].

4.3 Optical Flow

In this last section we present a non-medical application of non-linear registration, namely the estimation of optical flow [59]. Optical flow is the problem of determining the apparent motion in 2D images capturing a 3D scene. The motion is represented as a displacement field which defines the movement of pixels with 2D vectors. Thus, optical flow estimation is closely related to the problem of non-linear registration. Technically it is the same, however in optical flow we have to deal with issues such as occlusion, illumination changes, and shadows which all occur when capturing a real scene with camera images. Optical flow is studied for more than 30 years and a lot of progress has been made since the pioneering work of Horn and Schunck [59]. Currently, the most popular benchmark used for comparison of optical flow algorithms is the evaluation on the Middlebuary database[3] introduced by Baker et al. [4]. Most of our experiments on optical flow are performed on images from this database; we also use the color-coded visualization of flow fields (please see [4] for more details) which often is easier to interpret compared to displacement fields.

First, we will demonstrate an extension to our first-order MRF approach which deals with the automatic definition of the label space via uncertainty estimation. Then we demonstrate our latest results on optical flow estimation with our higher-order CRF approach combined with a multi-layer mesh representation.

4.3.1 Uncertainties

In our work on uncertainty estimation we were interested in the question whether it is possible to automatically define and adjust the label space respectively the discretization of the displacement search range. We should note that all experiments in this respect have been conducted on optical flow, while the actual method is applicable to other applications. The approach which we propose in [45] is based on the following idea: let us assume we have estimated the MAP labeling with respect to a particular set of labels; the labeling corresponds to an update on the control point displacements. Can we deduce from this solution how the search range for the next iteration should look like? Here, we were inspired by the work on uncertainty estimation by Kohli and Torr [80, 82].

Uncertainty estimation is the task of measuring how reliable a certain solution is. To be more precise, for every variable in the random field we are interested in measuring how certain (or confident) we are with respect to the label assigned to this variable. In [82] an efficient method based on dynamic graph cuts is presented which allows to exactly measure this uncertainty. Intuitively, it works as follows: imagine node i is assigned label x, i.e. the label in the MAP estimate. We consider this label as the optimal one with minimal energy. Now, in order to measure how confident we can be about this label, we need to have a look at the non-optimal labels, as well. In this context, it seems interesting to investigate the change of the energy in the case node i would have been assigned a non-optimal $k \in L$. A measure of confidence could then be defined as

$$\sigma_{i;x} = \frac{\exp(-\mu_{i;x})}{\sum_{k \in L} \exp(-\mu_{i;k})} \quad . \tag{4.13}$$

[3] `http://vision.middlebury.edu/flow/`

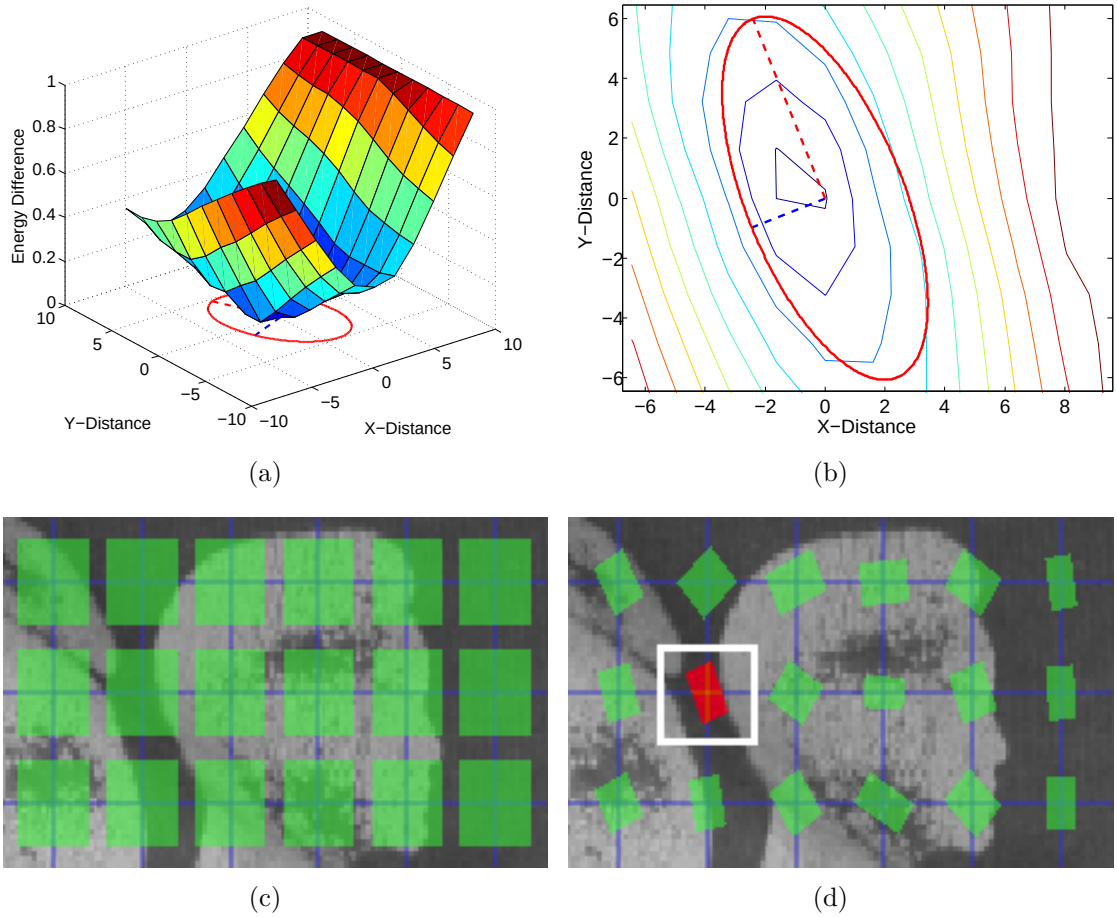


Figure 4.9: Uncertainty estimation for automatic label space adjustment. In (a) and (b) we show the min-marginal energy maps for an exemplary control point (the red one in (d)). In (c) the initial dense label spaces (in green) are shown, and in (d) the label spaces after local re-adjustment based on the uncertainty estimation.

Here, $\mu_{i;k}$ corresponds to the random field energy where node i is constrained to be assigned label k; such energies are also called the *min-marginals*. How the label constraints can be ensured is described [82] for the case of graph-cut optimization. The trick is to manipulate the capacities in the *st*-graph such that node i is guaranteed to be connected to the terminal representing label k. In [82] min-marginals are used to compute confidence maps in low-level vision applications. In image segmentation, the confidence map could for instance guide the user in which areas additional user input might be beneficial. In our case, these one-dimensional measures are not directly helpful, at least not for our actual objective: the adjustment of the label space.

Here, we do something different. Assume we have a dense discretization of the displacement space (cf. Figure 3.9(c)). Then we can assign the min-marginal energies to a location in Euclidean space. In fact, we can determine a min-marginal energy map for every control point in the deformation grid. From these energy maps we determine the covariance matrix which is then used to re-adjust the label space for the next iteration in terms of scale and orientation of the search range. This process is illustrated for an

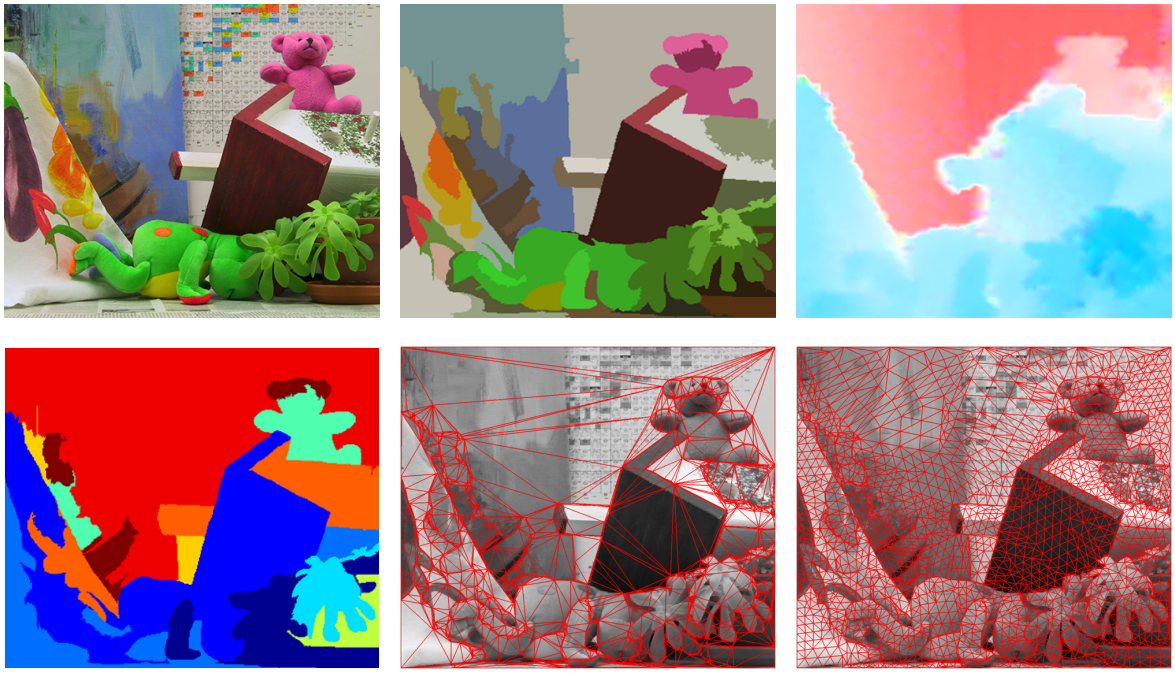


Figure 4.10: Multi-layer mesh construction for TriangleFlow. Top row shows from left to right the source image, its over-segmentation, the initial flow field. Bottom row shows the clustering result, the initial multi-layer mesh, and the final mesh after several refinements.

exemplary control point in Figure 4.9. Intuitively, as flatter the shape of the energy surface becomes in a particular direction as less reliable is the motion of the control along this direction. In other words, it would not make much difference from an energy point of view if this control point is displaced by one these non-optimal labels. Contrary, if the energy difference is large we are quite confident about the motion of the control point. In our experiments on optical flow [45], we could show that the fully automatic adjustment of the search space yields very accurate flow estimates without the need for empirically determine an appropriate refinement factor (cf. Section 3.4.5).

One interesting direction in this respect would also be to investigate the use of the original confidence maps for instance for visualizing registration uncertainty. In particular, in medical applications it could be very useful to have a visual feedback on the result after performing non-linear registration. A physician could make use of the information telling in which areas we are more confident and in which we are less confident about the alignment.

4.3.2 TriangleFlow

Our most recent work is on optical flow using the triangulation-based higher-order CRF model (cf. Section 3.4.4), hence we call this approach *TriangleFlow*. Remember, the higher-order model works as follows: the source image is covered by a triangulation mesh, and each triangle defines a local affine warp on the covered triangular sub-image. The warp is parameterized through displacements of the three triangle vertices. The regis-

tration energy is encoded on higher-order potential functions, i.e. triple-cliques for the likelihood (the similarity measure) and either triple- or quadruple-cliques for the geometric regularization (two different variants are introduced and compared in our work [38]). The great advantage of this model is that all energy terms are exact, meaning there is no approximation compared to the first-order MRF model. Additionally, triangulations are very flexible and can be adapted to actual image content.

For the optical flow experiments we employ the following strategy for the mesh construction: (i) given the two image frames, we first compute an initial flow field. In general, this could be done with any optical flow method. We use the higher-order model with a regular (single-layer) triangle mesh; (ii) we perform an over-segmentation of the source image; for every segment we estimate the closest affine warp (in a least-squares sense) from the initial motion field; (iii) we perform a clustering such that segments with similar affine motion fall into the same cluster; for each cluster we define a data-dependent triangulation mesh, i.e. a mesh which is aligned with object boundaries. Figure 4.10 shows the intermediate results of the single steps.

The multi-layer mesh model has great advantages compared to a regular content-unaware mesh. First, the actual motion of objects can be much better recovered with triangulations which are aligned to object boundaries. Second, the multi-layer mesh allows an explicit handling of occlusions which is important in case of optical flow. Actually, the layers can overlap in areas of occlusion and we explicitly evaluate which layer is on top of the others by considering the similarity measure within the overlap areas. The layer with highest similarity is the top layer, and the occluded areas can be discarded in the energy computation of lower layers. Some visual results are shown in Figure 4.11 where we compare the flow fields of a single-layer regular mesh approach and our multi-layer content-aware mesh approach. In all experiments, we use a coarse-to-fine strategy where the meshes are successively refined. We clearly see the improvement in particular at the motion boundaries. Very fine details are nicely recovered in the multi-layer results. A comprehensive quantitative evaluation can be found on the website of the database. Further experiments on other sequences are presented in our paper [38].

An interesting direction would be to integrate the whole process of triangulation and motion layer definition into the optimization. Flow-dependent mesh-refinement could further improve the results. A step beyond our current approach could allow for the definition of higher-order likelihoods with arbitrary shapes and without restrictions through the parametrization. We believe that in particular medical applications could benefit from such an approach. Ideally, we would like to have a transformation model which is capable of representing all the various motions which occur inside the human body. While bony structures will always have very limited deformation, soft tissue will behave totally different. Higher-order models have the expressional power to encode these different properties. A model which can be adapted to the specific anatomy, which is content-aware, and considers the locally varying physical properties of anatomical structures could be a key component for pushing non-linear registration a huge step forward. We believe that our random field models are an important first leap towards this future.

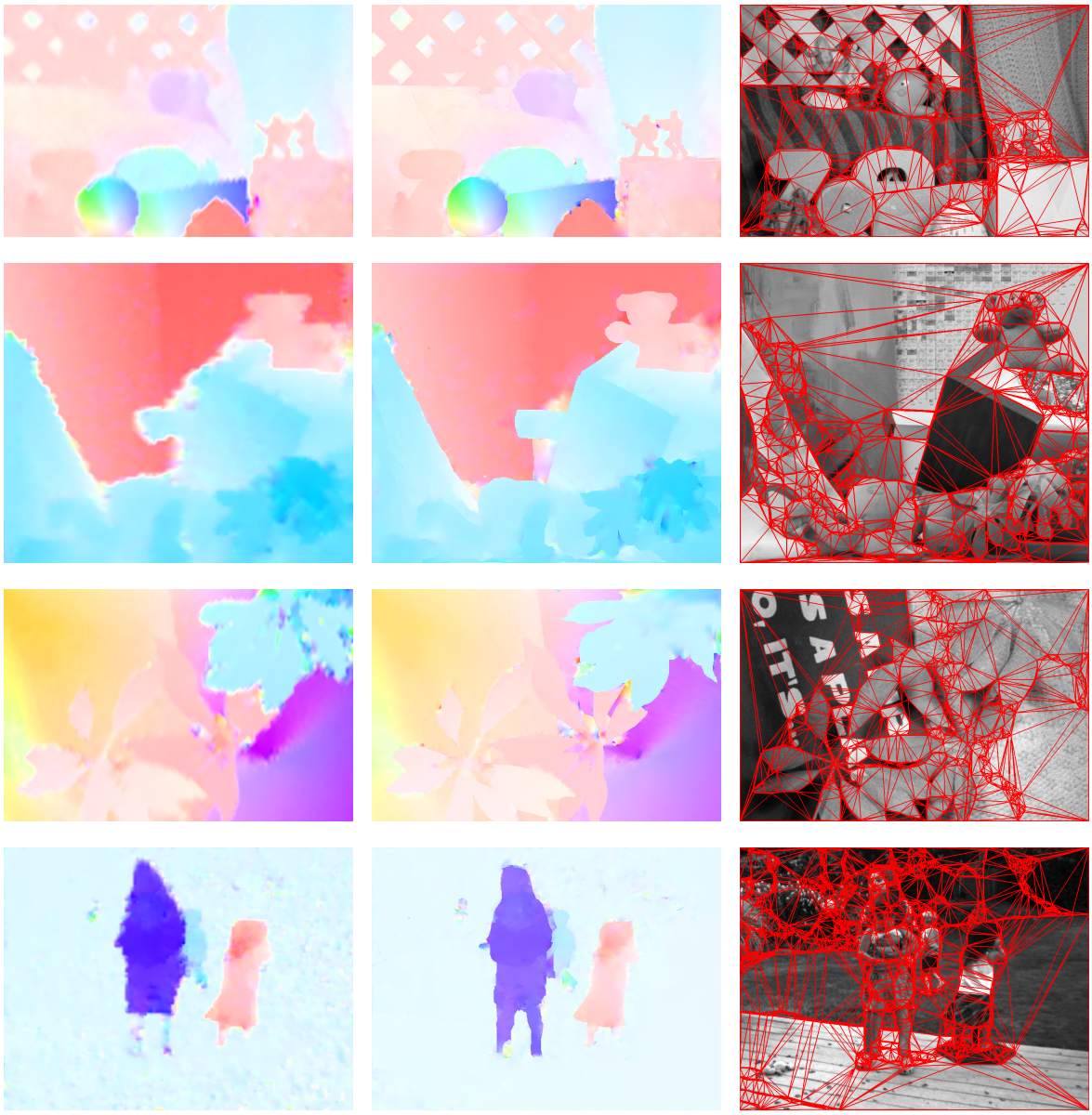


Figure 4.11: Optical flow estimation with our higher-order CRF model. On the left, we show the resulting color-encoded flow fields when using a regular single-layer mesh. In the middle the results when using our multi-layer mesh approach. The initial configurations of the multi-layer meshes overlaid on the source images are shown on the right.

LIST OF FIGURES

AUTHOR'S PUBLICATION LIST

[Atasoy et al., 2009] Atasoy, S., Glocker, B., Giannarou, S., Mateus, D., Meining, A., Yang, G.-Z., and Navab, N. (2009). Probabilistic Region Matching in Narrow-Band Endoscopy for Targeted Optical Biopsy. In *International Conference on Medical Image Computing and Computer Assisted Intervention*.

[Atasoy et al., 2008] Atasoy, S., Groher, M., Zikic, D., Glocker, B., Waggershauser, T., Pfister, M., and Navab, N. (2008). Real-Time Respiratory Motion Tracking: Roadmap Correction for Hepatic Artery Catheterizations. In *SPIE Medical Imaging*.

[Besbes et al., 2007] Besbes, A., Komodakis, N., Glocker, B., Tziritas, G., and Paragios, N. (2007). 4D Ventricular Segmentation and Wall Motion Estimation Using Efficient Discrete Optimization. In *International Symposium on Visual Computing*.

[Brieu et al., 2010] Brieu, N., Glocker, B., Navab, N., and Groher, M. (2010). MAP-MRF Optimal Partitioning for Dynamic Texture Segmentation of Thrombus in Time-Series Microscopic Images. In *Workshop Spatio Temporal Image Analysis for Longitudinal and Time-Series Image Data in conjunction with Medical Image Computing and Computer-Assisted Intervention*.

[Glocker et al., 2007a] Glocker, B., Buhmann, S., Kirchhoff, C., Mussack, T., Reiser, M., and Navab, N. (2007a). Towards a Computer Aided Diagnosis System for Colon Motility Dysfunctions. In *SPIE Medical Imaging*.

[Glocker et al., 2010a] Glocker, B., Heibel, H., Navab, N., Kohli, P., and Rother, C. (2010a). TriangleFlow: Optical Flow with Triangulation-based Higher-Order Likelihoods. In *European Conference on Computer Vision*.

[Glocker et al., 2009a] Glocker, B., Komodakis, N., Navab, N., Tziritas, G., and Paragios, N. (2009a). Dense Registration with Deformation Priors. In *Information Processing in Medical Imaging*.

[Glocker et al., 2007b] Glocker, B., Komodakis, N., Paragios, N., Glaser, C., Tziritas, G., and Navab, N. (2007b). Primal/Dual Linear Programming and Statistical Atlases for Cartilage Segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention*.

[Glocker et al., 2009b] Glocker, B., Komodakis, N., Paragios, N., and Navab, N. (2009b). Approximated Curvature Penalty in Non-rigid Registration using Pairwise MRFs. In *International Symposium on Visual Computing*.

[Glocker et al., 2010b] Glocker, B., Komodakis, N., Paragios, N., and Navab, N. (2010b). Non-rigid Registration using Discrete MRFs: Application to Thoracic CT Images. In *Workshop Evaluation of Methods for Pulmonary Image Registration in conjunction with Medical Image Computing and Computer-Assisted Intervention*.

[Glocker et al., 2007c] Glocker, B., Komodakis, N., Paragios, N., Tziritas, G., and Navab, N. (2007c). Inter and Intra-Modal Deformable Registration: Continuous Deformations Meet Efficient Optimal Linear Programming. In *Information Processing in Medical Imaging*.

[Glocker et al., 2008a] Glocker, B., Komodakis, N., Tziritas, G., Navab, N., and Paragios, N. (2008a). Dense Image Registration through MRFs and Efficient Linear Programming. *Medical Image Analysis*, 12(6):731–741.

[Glocker et al., 2008b] Glocker, B., Paragios, N., Komodakis, N., Tziritas, G., and Navab, N. (2008b). Optical Flow Estimation with Uncertainties through Dynamic MRFs. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[Glocker et al., 2009c] Glocker, B., Zikic, D., Komodakis, N., Paragios, N., and Navab, N. (2009c). Linear Image Registration through MRF Optimization. In *IEEE International Symposium on Biomedical Imaging*.

[Hansen et al., 2008] Hansen, M. S., Glocker, B., Navab, N., and Larsen, R. (2008). Adaptive Parametrization of Multivariate B-splines for Image Registration. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[Heibel et al., 2009] Heibel, T. H., Glocker, B., Groher, M., Paragios, N., Komodakis, N., and Navab, N. (2009). Discrete Tracking of Parametrized Curves. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[Heibel et al., 2010] Heibel, T. H., Glocker, B., Paragios, N., and Navab, N. (2010). Needle Tracking Through Higher-Order MRF Optimization. In *IEEE International Symposium on Biomedical Imaging*.

[Komodakis et al., 2009] Komodakis, N., Besbes, A., Glocker, B., and Paragios, N. (2009). Biomedical Image Analysis Using Markov Random Fields & Efficient Linear Programing. In *Internatinal Conference of the IEEE Engineering in Medicine and Biology Society*.

[Sotiras et al., 2009] Sotiras, A., Komodakis, N., Glocker, B., Deux, J.-F., and Paragios, N. (2009). Graphical Models and Deformable Diffeomorphic Population Registration Using Global and Local Metrics. In *International Conference on Medical Image Computing and Computer Assisted Intervention*.

[Sotiras et al., 2010] Sotiras, A., Ou, Y., Glocker, B., Davatzikos, C., and Paragios, N. (2010). Simultaneous Geometric - Iconic Registration. In *International Conference on Medical Image Computing and Computer Assisted Intervention*.

[Wachinger et al., 2009] Wachinger, C., Baumann, S., Zeltner, J., Glocker, B., , and Navab, N. (2009). Sphere Extraction in MR Images with Application to Whole-Body MRI. In *SPIE Medical Imaging*.

[Wachinger et al., 2008] Wachinger, C., Glocker, B., Zeltner, J., Paragios, N., Komodakis, N., Hansen, M. S., and Navab, N. (2008). Deformable Mosaicing for Whole-body MRI. In *International Conference on Medical Image Computing and Computer Assisted Intervention*.

[Zikic et al., 2008a] Zikic, D., Glocker, B., Hansen, M. S., Khamene, A., and Navab, N. (2008a). Construction of Statistical Shape Models from Minimal Deformations. In *Workshop Manifolds in Medical Imaging: Metrics, Learning and Beyond in conjunction with Medical Image Computing and Computer-Assisted Intervention*.

[Zikic et al., 2010a] Zikic, D., Glocker, B., Kutter, O., Groher, M., Komodakis, N., Kamen, A., Paragios, N., and Navab, N. (2010a). Linear Intensity-based Image Registration by Markov Random Fields and Discrete Optimization. *Medical Image Analysis (joint first authors)*, 14(4):550–562.

[Zikic et al., 2010b] Zikic, D., Glocker, B., Kutter, O., Groher, M., Komodakis, N., Khamene, A., Paragios, N., and Navab, N. (2010b). Markov Random Field Optimization for Intensity-based 2D-3D Registration. In *SPIE Medical Imaging*.

[Zikic et al., 2008b] Zikic, D., Hansen, M. S., Glocker, B., Khamene, A., Larsen, R., and Navab, N. (2008b). Computing Minimal Deformations: Application to Construction of Statistical Shape Models. In *IEEE Conference on Computer Vision and Pattern Recognition*.

REFERENCES

[1] Ahuja, R. K., Ergun, Ö., Orlin, J. B., and Punnen, A. P. A survey of very large-scale neighborhood search techniques. *Discrete Applied Mathematics 123*, 1–3 (2002), 75–102.

[2] Ali, A. M., Farag, A. A., and Gimel'farb, G. L. Optimizing Binary MRFs with Higher Order Cliques. In *European Conference on Computer Vision* (2008).

[3] Arun, K. S., Huang, T. S., and Blostein, S. D. Least-Squares Fitting of Two 3-D Point Sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence 9*, 5 (1987), 698–700.

[4] Baker, S., Scharstein, D., Lewis, J., Roth, S., Black, M. J., and Szeliski, R. A Database and Evaluation Methodology for Optical Flow. Tech. Rep. MSR-TR-2009-179, Microsoft Research, December 2009.

[5] Besag, J. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological) 36*, 2 (1974), 192–236.

[6] Besag, J. Statistical Analysis of Non-Lattice Data. *The Statistician 24*, 3 (1975), 179–195.

[7] Besag, J. On the statistical analysis of dirty images. *Journal of the Royal Statistical Society. Series B (Methodological) 48*, 3 (1986), 259–302.

[8] Besl, P. J., and McKay, H. D. A method for registration of 3-D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence 14*, 2 (1992), 239–256.

[9] Birchfield, S., and Tomasi, C. Multiway Cut for Stereo and Motion with Slanted Surfaces. In *IEEE International Conference on Computer Vision* (1999).

[10] Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer-Verlag, 2006.

[11] Blake, A., and Zisserman, A. *Visual Reconstruction*. MIT Press Cambridge, MA, 1987.

[12] Bookstein, F. L. Principal warps: thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence 11*, 6 (1989), 567–585.

[13] Boros, E., and Hammer, P. L. Pseudo-boolean optimization. *Discrete Applied Mathematics 123*, 1–3 (2002), 155–225.

[14] Boros, E., Hammer, P. L., and Sun, X. Network Flows and Minimization of Quadratic Pseudo-Boolean Functions. Tech. Rep. RRR 17-1991, RUTCOR, May 1991.

[15] Boykov, Y., and Jolly, M.-P. Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images. In *IEEE International Conference on Computer Vision* (2001).

[16] Boykov, Y., and Kolmogorov, V. An experimental comparison of min-cut/max- flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence 26*, 9 (2004), 1124–1137.

[17] Boykov, Y., Veksler, O., and Zabih, R. Markov random fields with efficient approximations. In *IEEE Conference on Computer Vision and Pattern Recognition* (1998).

[18] Boykov, Y., Veksler, O., and Zabih, R. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence 23*, 11 (2001), 1222–1239.

[19] Bronstein, M., Bronstein, A., Michel, F., and Paragios, N. Data Fusion through Cross-modality Metric Learning using Similarity-Sensitive Hashing. In *IEEE Conference on Computer Vision and Pattern Recognition* (2010).

[20] Cachier, P., Bardinet, E., Dormont, D., Pennec, X., and Ayache, N. Iconic feature based nonrigid registration: the PASHA algorithm. *Computer Vision and Image Understanding 89*, 2–3 (2003), 272–298.

[21] Cerny, V. Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. *Journal of Optimization Theory and Applications 45*, 1 (1985), 41–51.

[22] Chellappa, R., and Jain, A., Eds. *Markov Random Fields: Theory and Applications*. Academic Press Boston, 1993.

[23] Chou, P. B., and Brown, C. M. The Theory and Practice of Bayesian Image Labeling. *International Journal of Computer Vision 4*, 3 (1990), 185–210.

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

[25] Chung, A., Wells, W., Norbash, A., and Grimson, W. Multi-modal Image Registration by Minimizing Kullback-Leibler Distance. In *International Conference on Medical Image Computing and Computer Assisted Intervention* (2002).

[26] Cremers, D., and Grady, L. Statistical priors for efficient combinatorial optimization via graph cuts. In *European Conference on Computer Vision* (2006).

[27] Davatzikos, C., Prince, J., and Bryan, R. Image registration based on boundary mapping. *IEEE Transactions on Medical Imaging 15*, 1 (1996), 112–115.

[28] Faugeras, O. D., and Berthod, M. Improving Consistency and Reducing Ambiguity in Stochastic Labeling: An Optimization Approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence 3*, 4 (1981), 412–424.

[29] Feldmar, J., Declerck, J., Malandain, G., and Ayache, N. Extension of the ICP algorithm to nonrigid intensity-based registration of 3 D volumes. *Computer Vision and Image Understanding 66*, 2 (1997), 193–206.

[30] Felzenszwalb, P., and Huttenlocher, D. Efficient belief propagation for early vision. *International Journal of Computer Vision 70*, 1 (2006), 41–54.

[31] Fitzgibbon, A. Robust registration of 2D and 3D point sets. *Image and Vision Computing 21*, 13–14 (2003), 1145–1153.

[32] Ford, L., and Fulkerson, D. Maximal flow through a network. *Canadian Journal of Mathematics 8*, 3 (1956), 399–404.

[33] Freedman, D., and Drineas, P. Energy Minimization via Graph Cuts: Settling What is Possible. In *IEEE Conference on Computer Vision and Pattern Recognition* (2005).

[34] Freeman, W., Pasztor, E., and Carmichael, O. Learning low-level vision. *International Journal of Computer Vision 40*, 1 (2000), 25–47.

[35] Frey, B. J., and MacKay, D. J. C. A Revolution: Belief Propagation in Graphs with Cycles. *Advances in Neural Information Processing Systems 10* (1997), 479–485.

[36] Geman, S., and Geman, D. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence 6*, 6 (1984), 721–741.

[37] Geman, S., and Graffigne, C. Markov random field image models and their applications to computer vision. In *Proceedings of the International Congress of Mathematicians* (1986), pp. 1496–1517.

[38] Glocker, B., Heibel, H., Navab, N., Kohli, P., and Rother, C. TriangleFlow: Optical Flow with Triangulation-based Higher-Order Likelihoods. In *European Conference on Computer Vision* (2010).

[39] Glocker, B., Komodakis, N., Navab, N., Tziritas, G., and Paragios, N. Dense Registration with Deformation Priors. In *Information Processing in Medical Imaging* (2009).

[40] Glocker, B., Komodakis, N., Paragios, N., Glaser, C., Tziritas, G., and Navab, N. Primal/Dual Linear Programming and Statistical Atlases for Cartilage Segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention* (2007).

[41] Glocker, B., Komodakis, N., Paragios, N., and Navab, N. Approximated Curvature Penalty in Non-rigid Registration using Pairwise MRFs. In *International Symposium on Visual Computing* (2009).

[42] Glocker, B., Komodakis, N., Paragios, N., and Navab, N. Non-rigid Registration using Discrete MRFs: Application to Thoracic CT Images. In *Workshop Evaluation of Methods for Pulmonary Image Registration in conjunction with Medical Image Computing and Computer-Assisted Intervention* (2010).

[43] Glocker, B., Komodakis, N., Paragios, N., Tziritas, G., and Navab, N. Inter and Intra-Modal Deformable Registration: Continuous Deformations Meet Efficient Optimal Linear Programming. In *Information Processing in Medical Imaging* (2007).

[44] Glocker, B., Komodakis, N., Tziritas, G., Navab, N., and Paragios, N. Dense Image Registration through MRFs and Efficient Linear Programming. *Medical Image Analysis 12*, 6 (2008), 731–741.

[45] Glocker, B., Paragios, N., Komodakis, N., Tziritas, G., and Navab, N. Optical Flow Estimation with Uncertainties through Dynamic MRFs. In *IEEE Conference on Computer Vision and Pattern Recognition* (2008).

[46] Glocker, B., Zikic, D., Komodakis, N., Paragios, N., and Navab, N. Linear Image Registration through MRF Optimization. In *IEEE International Symposium on Biomedical Imaging* (2009).

[47] Goldberg, A., and Tarjan, R. A new approach to the maximum-flow problem. *Journal of the ACM 35*, 4 (1988), 921–940.

[48] Greig, D., Porteous, B., and Seheult, A. Exact maximum a posteriori estimation for binary images. *Journal of the Royal Statistical Society. Series B (Methodological) 51*, 2 (1989), 271–279.

[49] Hajnal, J., Hill, D. L. G., and Hawkes, D. J., Eds. *Medical Image Registration*. CRC Press, 2001.

[50] Hammer, P., Hansen, P., and Simeone, B. Roof duality, complementation and persistency in quadratic 0–1 optimization. *Mathematical Programming 28*, 2 (1984), 121–155.

[51] Hammersley, J., and Clifford, P. Markov Fields on Finite Graphs and Lattices. Unpublished Manuscript, 1971.

[52] Hansen, M., Glocker, B., Navab, N., and Larsen, R. Adaptive Parametrization of Multivariate B-splines for Image Registration. In *IEEE Conference on Computer Vision and Pattern Recognition* (2008).

[53] Hartley, R., and Zisserman, A. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2003.

[54] He, Y., He, Y., Hamza, A., and Krim, H. A generalized divergence measure for robust image registration. *IEEE Transactions on Signal Processing 51*, 5 (2003), 1211–1220.

[55] Hellier, P., and Barillot, C. Coupling dense and landmark-based approaches for nonrigid registration. *IEEE Transactions on Medical Imaging 22*, 2 (2003), 217–227.

[56] Hermosillo, G., Chefd'Hotel, C., and Faugeras, O. Variational methods for multimodal image matching. *International Journal of Computer Vision 50*, 3 (2002), 329–343.

[57] Holden, M. A review of geometric transformations for nonrigid body registration. *IEEE Transactions on Medical Imaging 27*, 1 (2008), 111–128.

[58] Horn, B. *Robot Vision*. MIT Press Cambridge, MA, 1986.

[59] Horn, B., and Schunck, B. Determining optical flow. *Artificial intelligence 17*, 1-3 (1981), 185–203.

[60] Hummel, R. A., and Zucker, S. W. On the Foundations of Relaxation Labeling Processes. *PAMI 5*, 3 (1983), 267–287.

[61] Ishikawa, H. *Global Optimization Using Embedded Graphs*. PhD thesis, New York University, 2000.

[62] Ishikawa, H. Exact optimization for Markov random fields with convex priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence 25*, 10 (2003), 1333–1336.

[63] Ishikawa, H. Higher-order clique reduction in binary graph cut. In *IEEE Conference on Computer Vision and Pattern Recognition* (2009).

[64] Ishikawa, H. Higher-Order Gradient Descent by Fusion-Move Graph Cut. In *IEEE International Conference on Computer Vision* (2009).

[65] Ishikawa, H. Transformation of General Binary MRF Minimization to the First Order Case. *IEEE Transactions on Pattern Analysis and Machine Intelligence PrePrint* (2010).

[66] Johnson, H., and Christensen, G. Consistent landmark and intensity-based image registration. *IEEE Transactions on Medical Imaging 21*, 5 (2002), 450–461.

[67] Joshi, S., Davis, B., Jomier, M., and Gerig, G. Unbiased diffeomorphic atlas construction for computational anatomy. *NeuroImage 23* (2004), S151–S160.

[68] Joshi, S., and Miller, M. Landmark matching via large deformation diffeomorphisms. *IEEE Transactions on Image Processing 9*, 8 (2000), 1357–1370.

[69] Kindermann, R., and Snell, J. *Markov random fields and their applications*. American Mathematical Society, 1980.

[70] Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. Optimization by Simulated Annealing. *Science 220*, 4598 (1983), 671–680.

[71] Kittler, J., and Illingworth, J. Relaxation Labelling Algorithms – A Review. *Image and Vision Computing 3*, 4 (1985), 206–216.

[72] Klein, S., Pluim, J. P. W., Staring, M., and Viergever, M. Adaptive stochastic gradient descent optimisation for image registration. *International Journal of Computer Vision 81*, 3 (2009), 227–239.

[73] Klein, S., Staring, M., Murphy, K., Viergever, M., and Pluim, J. elastix: A Toolbox for Intensity-based Medical Image Registration. *IEEE Transactions on Medical Imaging 29*, 1 (2010), 196–205.

[74] Klein, S., Staring, M., and Pluim, J. P. W. Evaluation of Optimization Methods for Nonrigid Medical Image Registration Using Mutual Information and B-Splines. *IEEE Transactions on Image Processing 16*, 12 (2007), 2879–2890.

[75] Kohli, P. *Minimizing Dynamic and Higher Order Energy Functions using Graph Cuts*. PhD thesis, Oxford Brookes University, 2007.

[76] Kohli, P., Kumar, M. P., and Torr, P. H. S. P3 & Beyond: Solving Energies with Higher Order Cliques. In *IEEE Conference on Computer Vision and Pattern Recognition* (2007).

[77] Kohli, P., Ladický, L., and Torr, P. H. S. Robust higher order potentials for enforcing label consistency. In *IEEE Conference on Computer Vision and Pattern Recognition* (2008).

[78] Kohli, P., Ladický, L., and Torr, P. H. S. Robust Higher Order Potentials for Enforcing Label Consistency. *International Journal of Computer Vision 82*, 3 (2009), 302–324.

[79] Kohli, P., and Torr, P. H. S. Efficiently solving dynamic Markov random fields using graph cuts. In *IEEE International Conference on Computer Vision* (2005).

[80] Kohli, P., and Torr, P. H. S. Measuring Uncertainty in Graph Cut Solutions. In *European Conference on Computer Vision* (2006).

[81] Kohli, P., and Torr, P. H. S. Dynamic Graph Cuts for Efficient Inference in Markov Random Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence 29*, 12 (2007), 2079–2088.

[82] Kohli, P., and Torr, P. H. S. Measuring uncertainty in graph cut solutions. *Computer Vision and Image Understanding 112*, 1 (2008), 30–38.

[83] Kolmogorov, V. Convergent Tree-Reweighted Message Passing for Energy Minimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence 28*, 10 (2006), 1568–1583.

[84] Kolmogorov, V., and Rother, C. Comparison of Energy Minimization Algorithms for Highly Connected Graphs. In *European Conference on Computer Vision* (2006).

[85] Kolmogorov, V., and Rother, C. Minimizing Nonsubmodular Functions with Graph Cuts – A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence 29*, 7 (2007), 1274–1279.

[86] Kolmogorov, V., and Zabih, R. What energy functions can be minimized via graph cuts? *IEEE Transactions on Pattern Analysis and Machine Intelligence 26*, 2 (2004), 147–159.

[87] Komodakis, N. *Optimization Algorithms for Discrete Markov Random Fields, with Applications to Computer Vision.* PhD thesis, University of Crete, 2006.

[88] Komodakis, N., and Paragios, N. Beyond pairwise energies: Efficient optimization for higher-order MRFs. In *IEEE Conference on Computer Vision and Pattern Recognition* (2009).

[89] Komodakis, N., Paragios, N., and Tziritas, G. MRF Optimization via Dual Decomposition: Message-Passing Revisited. In *IEEE International Conference on Computer Vision* (2007).

[90] Komodakis, N., and Tziritas, G. A new framework for approximate labeling via graph cuts. In *IEEE International Conference on Computer Vision* (2005).

[91] Komodakis, N., and Tziritas, G. Approximate Labeling via Graph Cuts Based on Linear Programming. *IEEE Transactions on Pattern Analysis and Machine Intelligence 29*, 8 (2007), 1436–1453.

[92] Komodakis, N., Tziritas, G., and Paragios, N. Fast, Approximately Optimal Solutions for Single and Dynamic MRFs. In *IEEE Conference on Computer Vision and Pattern Recognition* (2007).

[93] Komodakis, N., Tziritas, G., and Paragios, N. Performance vs computational efficiency for optimizing single and dynamic MRFs: Setting the state of the art with primal-dual strategies. *Computer Vision and Image Understanding 112*, 1 (2008), 14–29.

[94] Kwon, D., Lee, K., Yun, I., and Lee, S. Nonrigid Image Registration Using Dynamic Higher-Order MRF Model. In *European Conference on Computer Vision* (2008).

[95] Lan, X., Roth, S., Huttenlocher, D., and Black, M. Efficient belief propagation with learned higher-order markov random fields. In *European Conference on Computer Vision* (2006).

[96] Lee, D., Hofmann, M., Steinke, F., Altun, Y., Cahill, N. D., and Scholkopf, B. Learning similarity measure for multi-modal 3D image registration. In *IEEE Conference on Computer Vision and Pattern Recognition* (2009).

[97] Lee, K. J., Kwon, D., Yun, I. D., and Lee, S. U. Deformable 3D Volume Registration Using Efficient MRFs Model with Decomposed Nodes. In *British Machine Vision Conference* (2008).

[98] Lee, S., Wolberg, G., Chwa, K., and Shin, S. Image metamorphosis with scattered feature constraints. *IEEE Transactions on Visualization and Computer Graphics 2*, 4 (1996), 337–354.

[99] Lempitsky, V., Roth, S., and Rother, C. FusionFlow: Discrete-continuous optimization for optical flow estimation. In *IEEE Conference on Computer Vision and Pattern Recognition* (2008).

[100] Lempitsky, V., Rother, C., and Blake, A. Logcut – efficient graph cut optimization for markov random fields. In *IEEE International Conference on Computer Vision* (2007).

[101] Lempitsky, V., Rother, C., Roth, S., and Blake, A. Fusion Moves for Markov Random Field Optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence 32*, 8 (2010), 1392–1405.

[102] Li, S. Z. *Markov Random Field Modeling in Image Analysis*. Springer-Verlag, 2009.

[103] Maes, F., Collignon, A., Vandermeulen, D., Marchal, G., and Suetens, P. Multimodality image registration by maximization of mutual information. *IEEE Transactions on Medical Imaging 16*, 2 (1997), 187–198.

[104] Mahamud, S. Comparing Belief Propagation and Graph Cuts for Novelty Detection. In *IEEE Conference on Computer Vision and Pattern Recognition* (2006).

[105] Maintz, J. B. A., and Viergever, M. A. A survey of medical image registration. *Medical Image Analysis 2*, 1 (1998), 1–36.

[106] Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., Teller, E., et al. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics 21*, 6 (1953), 1087–1092.

[107] Modersitziki, J. *Numerical Methods for Image Registration*. Oxford University Press, 2004.

[108] Nemhauser, G., and Wolsey, L. *Integer and combinatorial optimization*. Wiley New York, 1999.

[109] Otten, R. H. J. M., and van Ginneken, L. P. P. P. *The Annealing Algorithm*. Kluwer Academic Publishers, 1989.

[110] Ou, Y., and Davatzikos, C. DRAMMS: deformable registration via attribute matching and mutual-saliency weighting. In *Information Processing in Medical Imaging* (2009).

[111] Papadimitriou, C. H., and Steiglitz, K. *Combinatorial Optimization*. Dover Publications, Inc., 1998.

[112] Paragios, N., Ayache, N., and Duncan, J., Eds. *Biomedical Image Analysis: Methodologies and Applications*. Springer Verlag, 2010.

[113] Paragios, N., Chen, Y., and Faugeras, O., Eds. *Handbook of Mathematical Models in Computer Vision*. Springer Verlag, 2005.

[114] Pauly, O., Padoy, N., Poppert, H., Esposito, L., Eckstein, H.-H., and Navab, N. Towards Application-specific Multi-modal Similarity Measures: A Regression Approach. In *Workshop Probabilistic Models in Medical Image Analysis in conjunction with Medical Image Computing and Computer-Assisted Intervention* (2009).

[115] Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[116] Picard, J., and Ratliff, H. Minimum cuts and related problems. *Networks 5*, 4 (1975), 357–370.

[117] Potetz, B., and Lee, T. S. Efficient belief propagation for higher-order cliques using linear constraint nodes. *Computer Vision and Image Understanding 112*, 1 (2008), 39–54.

[118] Press, W., Teukolsky, S., Vetterling, W., and Flannery, B. *Numerical Recipes in C*. Cambridge University Press, 1993.

[119] Raj, A., Singh, G., and Zabih, R. MRFs for MRIs: Bayesian Reconstruction of MR Images via Graph Cuts. In *IEEE Conference on Computer Vision and Pattern Recognition* (2006).

[120] Raj, A., and Zabih, R. A Graph Cut Algorithm for Generalized Image Deconvolution. In *IEEE International Conference on Computer Vision* (2005).

[121] Rajwade, A., Banerjee, A., and Rangarajan, A. A New Method of Probability Density Estimation with Application to Mutual Information based Image Registration. In *IEEE Conference on Computer Vision and Pattern Recognition* (2006).

[122] Roche, A., Malandain, G., Pennec, X., and Ayache, N. The Correlation Ratio as a New Similarity Measure for Multimodal Image Registration. In *International Conference on Medical Image Computing and Computer Assisted Intervention* (1998).

[123] Rogelj, P., Kovai, S., and Gee, J. Point similarity measures for non-rigid registration of multi-modal data. *Computer Vision and Image Understanding 92*, 1 (2003), 112–140.

[124] Rohr, K., Stiehl, H., Sprengel, R., Buzug, T., Weese, J., and Kuhn, M. Landmark-based elastic registration using approximating thin-plate splines. *IEEE Transactions on Medical Imaging 20*, 6 (2001), 526–534.

[125] Rosenfeld, A., Hummel, R., and Zucker, S. Scene labeling by relaxation operations. *IEEE Transactions on Systems, Man and Cybernetics 6*, 6 (1976), 420–433.

[126] Rother, C., Kohli, P., Feng, W., and Jia, J. Minimizing sparse higher order energy functions of discrete variables. In *IEEE Conference on Computer Vision and Pattern Recognition* (2009).

[127] Rother, C., Kolmogorov, V., and Blake, A. Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM SIGGRAPH 23*, 3 (2004), 309–314.

[128] Rother, C., Kolmogorov, V., Lempitsky, V., and Szummer, M. Optimizing Binary MRFs via Extended Roof Duality. In *IEEE Conference on Computer Vision and Pattern Recognition* (2007).

[129] Rother, C., Kumar, S., Kolmogorov, V., and Blake, A. Digital tapestry [automatic image synthesis]. In *IEEE Conference on Computer Vision and Pattern Recognition* (2005).

[130] Roy, S., and Cox, I. J. A maximum-flow formulation of the N-camera stereo correspondence problem. In *IEEE International Conference on Computer Vision* (1998).

[131] Roy, S., and Govindu, V. MRF Solutions for Probabilistic Optical Flow Formulations. In *International Conference on Pattern Recognition* (2000).

[132] Rueckert, D., Aljabar, P., Heckemann, R., Hajnal, J., and Hammers, A. Diffeomorphic registration using B-splines. In *International Conference on Medical Image Computing and Computer Assisted Intervention* (2006).

[133] Rueckert, D., Sonoda, L. I., Hayes, C., Hill, D. L. G., Leach, M. O., and Hawkes, D. J. Nonrigid registration using free-form deformations: application to breast MR images. *IEEE Transactions on Medical Imaging 18*, 8 (1999), 712–721.

[134] Scharstein, D., and Szeliski, R. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision 47*, 1–3 (2002), 7–42.

[135] Sederberg, T. W., and Parry, S. R. Free-form deformation of solid geometric models. *ACM SIGGRAPH 20*, 4 (1986), 151–160.

[136] Shekhovtsov, A., Kovtun, I., and Hlavac, V. Efficient MRF Deformation Model for Non-Rigid Image Matching. In *IEEE Conference on Computer Vision and Pattern Recognition* (2007).

[137] Shekhovtsov, A., Kovtun, I., and Hlavac, V. Efficient MRF deformation model for non-rigid image matching. *Computer Vision and Image Understanding 112*, 1 (2008), 91–99.

[138] Shen, D., and Davatzikos, C. HAMMER: hierarchical attribute matching mechanism for elastic registration. *IEEE Transactions on Medical Imaging 21*, 11 (2002), 1421–1439.

[139] Smyth, P. Belief networks, hidden Markov models, and Markov random fields: a unifying view. *Pattern Recognition Letters 18*, 11–13 (1997), 1261–1268.

[140] Sotiras, A., Ou, Y., Glocker, B., Davatzikos, C., and Paragios, N. Simultaneous Geometric - Iconic Registration. In *International Conference on Medical Image Computing and Computer Assisted Intervention* (2010).

[141] Studholme, C., Hill, D., and Hawkes, D. An overlap invariant entropy measure of 3D medical image alignment. *Pattern Recognition 32* (1999), 71–86.

[142] Szeliski, R. Image Alignment and Stitching: A Tutorial. *Foundations and Trends in Computer Graphics and Vision 2*, 1 (2006), 1–104.

[143] Szeliski, R., Zabih, R., Scharstein, D., Veksler, O., Kolmogorov, V., Agarwala, A., Tappen, M., and Rother, C. A Comparative Study of Energy Minimization Methods for Markov Random Fields with Smoothness-Based Priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence 30*, 6 (2008), 1068–1080.

[144] Tang, T., and Chung, A. Non-rigid image registration using graph-cuts. In *International Conference on Medical Image Computing and Computer Assisted Intervention* (2007).

[145] Tappen, M. F., and Freeman, W. T. Comparison of graph cuts with belief propagation for stereo, using identical MRF parameters. In *IEEE International Conference on Computer Vision* (2003).

[146] Thirion, J. Image matching as a diffusion process: an analogy with Maxwell's demons. *Medical Image Analysis 2*, 3 (1998), 243–260.

[147] Tikhonov, A., Goncharsky, A., and Bloch, M. *Ill-posed problems in the natural sciences*. Mir Moscow, 1987.

[148] Umeyama, S. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence 13*, 4 (1991), 376–380.

[149] Veksler, O. *Efficient Graph-based Energy Minimization Methods in Computer Vision*. PhD thesis, Cornell University, 1999.

[150] Veksler, O. Graph Cut Based Optimization for MRFs with Truncated Convex Priors. In *IEEE Conference on Computer Vision and Pattern Recognition* (2007).

[151] Vercauteren, T., Pennec, X., Perchant, A., and Ayache, N. Diffeomorphic demons: Efficient non-parametric image registration. *NeuroImage 45*, 1 (2009), S61–S72.

[152] Viola, P., and Wells, W. Alignment by Maximization of Mutual Information. *International Journal of Computer Vision 24*, 2 (1997), 137–154.

[153] Wachinger, C., Glocker, B., Zeltner, J., Paragios, N., Komodakis, N., Hansen, M. S., and Navab, N. Deformable Mosaicing for Whole-body MRI. In *International Conference on Medical Image Computing and Computer Assisted Intervention* (2008).

[154] Wainwright, M., Jaakkola, T., and Willsky, A. Tree consistency and bounds on the performance of the max-product algorithm and its generalizations. *Statistics and Computing 14*, 2 (2004), 143–166.

[155] Wainwright, M. J., Jaakkola, T. S., and Willsky, A. S. MAP estimation via agreement on trees: message-passing and linear programming. *IEEE Transactions on Information Theory 51*, 11 (2005), 3697–3717.

[156] Weiss, Y., and Freeman, W. T. On the optimality of solutions of the max-product belief-propagation algorithm in arbitrary graphs. *IEEE Transactions on Information Theory 47*, 2 (2001), 736–744.

[157] Woodford, O., Torr, P., Reid, I., and Fitzgibbon, A. Global Stereo Reconstruction under Second-Order Smoothness Priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence 31*, 12 (2009), 2115–2128.

[158] Woodford, O. J., Torr, P. H. S., Reid, I. D., and Fitzgibbon, A. W. Global stereo reconstruction under second order smoothness priors. In *IEEE Conference on Computer Vision and Pattern Recognition* (2008).

[159] Woods, J. Two-dimensional discrete Markovian fields. *IEEE Transactions on Information Theory 18*, 2 (1972), 232–240.

[160] Yedidia, J., Freeman, W., and Weiss, Y. Generalized belief propagation. *Advances in Neural Information Processing Systems 13* (2000), 689–695.

[161] Yedidia, J. S., Freeman, W. T., , and Weiss, Y. Understanding Belief Propagation and its Generalizations. Tech. Rep. TR-2001-22, Mitsubishi Electric Research Laboratories, January 2002.

[162] Zikic, D., Glocker, B., Hansen, M. S., Khamene, A., and Navab, N. Construction of Statistical Shape Models from Minimal Deformations. In *Workshop Manifolds in Medical Imaging: Metrics, Learning and Beyond in conjunction with Medical Image Computing and Computer-Assisted Intervention* (2008).

[163] Zikic, D., Glocker, B., Kutter, O., Groher, M., Komodakis, N., Kamen, A., Paragios, N., and Navab, N. Linear Intensity-based Image Registration by Markov Random Fields and Discrete Optimization. *Medical Image Analysis (joint first authors) 14*, 4 (2010), 550–562.

[164] Zikic, D., Glocker, B., Kutter, O., Groher, M., Komodakis, N., Khamene, A., Paragios, N., and Navab, N. Markov Random Field Optimization for Intensity-based 2D-3D Registration. In *SPIE Medical Imaging* (2010).

[165] Zikic, D., Kamen, A., and Navab, N. Unifying Characterization of Deformable Registration Methods Based on the Inherent Parameterization. In *International Workshop on Biomedical Image Registration* (2010).

[166] Zitova, B., and Flusser, J. Image registration methods: a survey. *Image and Vision Computing 21*, 11 (2003), 977–1000.

[167] Zollei, L., Fisher, J., and Wells, W. *Handbook of Mathematical Models in Computer Vision*. Springer-Verlag, 2005, ch. An Introduction to Statistical Methods of Medical Image Registration.