Technische Universität München

Fakultät für Informatik
Lehrstuhl für Bildverarbeitung und Mustererkennung

Convex Semantic Priors for Continuous Multi-Label Optimization

Mohamed Souiai

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. Björn Menze
Prüfer der Dissertation:	1. Univ.-Prof. Dr. Daniel Cremers
	2. Univ.-Prof. Dr. Bastian Goldlücke Universität Konstanz

Die Dissertation wurde am 03. Januar 2017 bei der Technischen Universität München eingereicht und durch die Fakultät für Informatik am 21. Mai angenommen.

Contents

Abstract

Traditional image segmentation algorithms have merely focused on separating regions based on homogeneity of color or texture. More recent methods aim at incorporating high-level knowledge such as semantics or rigid body motion into what is often called class-based image segmentation. This thesis considers the problems of continuous semantic image segmentation and joint motion estimation and segmentation. The aim is to impose prior knowledge about which set of labels are likely to co-occur or how label transitions should take place between different rigid body motions.

Specifically, we propose to incorporate co-occurrence and hierarchical priors into continuous multi-label image segmentation. Similar to how humans learn to analyze a scene, these priors can be learned from co-occurrence statistics or drawn from natural label hierarchies. The latter prior is realized by assigning each label to a scene label such as 'indoor', 'outdoor', 'nature' or 'urban'. Arising from a continuous convex relaxation formulation, the resulting algorithms are, in contrast to discrete methods, not prone to metrification artifacts. Unlike sequential methods based on alpha-expansion, our optimization problems can be solved globally optimal using GPU-accelerated efficient primal-dual algorithms. Qualitative and quantitative results demonstrate the effectiveness of these priors in terms of the resulting labeling.

Furthermore, we propose a novel joint registration and segmentation approach applied to RGB-D images. Instead of assuming the scene to be composed of a number of independent rigidly-moving parts, we allow a smooth label transition in order to capture non-rigid deformations at transitions between the rigid parts of the scene. By doing so our method is able to infer the underlying motion estimates more accurately than state of the art works. At the same time, it provides a meaningful segmentation of the scene based on these motion cues.

Finally, we introduce a novel approach to improve the integrality of the solution of convex relaxed multi-labeling problems. Despite their enormous success in solving hard combinatorial problems, convex relaxation methods often suffer from the fact that the computed indicator functions are far from binary. Subsequent heuristic binarization may even substantially degrade the quality of computed solutions. To this end, we incorporate the entropy of the objective variable as a measure of the relaxation tightness. We use difference of convex function (DC) programming as an efficient and provably convergent solver for this convex-concave minimization problem. We evaluate this approach on three prominent computer vision problems: multi-label inpainting, image segmentation and spatio-temporal multi-view reconstruction. These experiments show that our approach consistently yields better solutions with respect to the original integer optimization problem.

Acknowledgements

Firstly, I would like to thank my advisor Prof. Daniel Cremers for his excellent supervision and for giving me the opportunity to work at such an outstanding group. I was lucky to have worked with my brilliant colleagues Evgeny Strekalovskiy, Claudia Nieuwenhuis, Youngwook Kee, Martin Oswald, Mariano Jaimez Tarifa, Julia Diebold, Christian Kerl and Vladimir Golkov. Without them this thesis wouldnt have been possible. I would also thank Sabine Wagner for helping me with all organizational things and being such a comforting person to have around. I would also like to thank my colleagues Thomas Möllenhoff, Michael Möller, Matthias Vestner, Thomas Windheuser and Emanuel Laude for fruitful discussions. During my thesis, I had the pleasure to share an office with Jürgen Sturm, Michael Möller and Vladimir Golkov who where very inspiring people on a personal and a professional level. I am grateful for Tainá Luersen, Matthias Vestner, Frank Schmidt, Thomas Möllenhoff and Vladimir Golkov for proofreading parts of my thesis. Finally, I would like to thank my family for its unconditional support.

Part I

Introduction and Overview

Chapter 1

Introduction

The core challenge in computer vision is to implement basic human visual perception on digital computers. This includes depth estimation, object recognition, motion estimation, and image segmentation. These disciplines are the building blocks of scene understanding which is one of the most important objectives in computer vision.

Of particular importance is the task of image segmentation which is the process of dissecting an image into its constituent parts. The image segmentation problem is a trivial task to human perception. Nevertheless, it remains a highly challenging problem in computer vision and despite it being a classical problem, it is still one of the most trending topics in the computer vision and image processing community today.

Subdividing an image into segments can be considered an instance of the broader class of multi-labeling problems. Assigning labels is a crucial task in machine vision because labels help explain input data and therefore reduce its complexity for further processing and understanding the present scene. Input data can be pixelwise color information from a camera, range information from a depth sensor or data from magnetic resonance imaging.

The applications areas are numerous since multi-labeling is useful whenever human vision is required. One trending application area is autonomous driving where the car needs to detect the road, persons, trees and other cars to avoid potential collisions. Another major application is the field of robotics, where a robot depends on understanding its environment to interact with it, for example to bring milk from the fridge it needs to recognize these as such.

This thesis deals with multi-labeling in various settings, e.g. semantic image segmentation, motion segmentation and 3D reconstruction. Depending on the application, labels can either describe the world geometrically, e.g. depth labels and object-void labels in 3D reconstruction or describe context information by assigning semantic labels such as 'car', 'road', 'tree'. Different scenarios of labeling applications are illustrated in Figure 1.1. Images (a) and (c) illustrate semantic labels in contrast to the labels in (b) and (d) which are of geometric nature and encode depth values and voxel occupancy respectively.

Occurring semantic labels differ depending on the scene type such as nature, urban, indoor or outdoor. Consequently, prior contextual information and the probability of certain label configurations are crucial for assigning the right labels. For this reason, modern multi-labeling algorithms aim at incorporating prior knowledge about the likelihood of a set of labels to occur. This knowledge can be based on learning co-occurrence statistics from ground truth data or on natural label hierarchies.

Another cornerstone to understanding a scene is estimating the motion of objects in a dynamic environment. For example, autonomous cars are reliant on estimating the motion of objects on a road in order to react correctly, e.g. knowing the motion of a crossing pedestrian is important to stop the car or not.

This thesis investigates incorporating semantic priors into variational multi-label algorithms. Additionally, we improve on the quality of the solutions of variational multi-labeling algorithms by jointly minimizing an entropy term. Finally, the problem of jointly segmenting an image and estimating the underlying motions of each part is tackled.

Semantic Priors

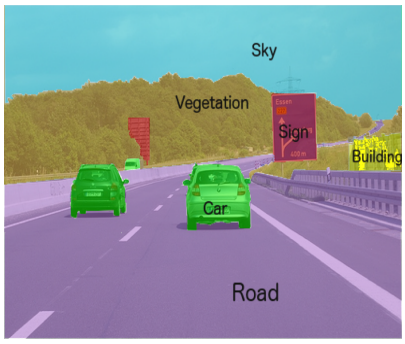
The algorithm we introduce in [10] allows for the integration of prior knowledge about what labels are likely to co-occur in the same image. For instance, in a scene where a person is riding an animal, our approach would facilitate a computer in labeling the animal as a horse rather than a moose because persons have a higher likelihood of being in the same image as a horse.

In [8], we devise a more principled way of incorporating semantic priors. We propose a joint approach for segmentation, object recognition, and scene understanding, which is formulated within a single multi-label variational optimization approach. Specifically, the central idea of this approach is to incorporate natural label hierarchies (i.e., assigning each image label to a scene label such as "indoor", "outdoor", "nature", "urban", etc.) in continuous multi-label segmentation. These natural label hierarchies are beneficial to the labeling process as they facilitate a computer, for instance, being able to decide the appropriate label of an animal that is located in an urban scene (e.g. ,downtown area) would be a dog or cat rather than a lion.

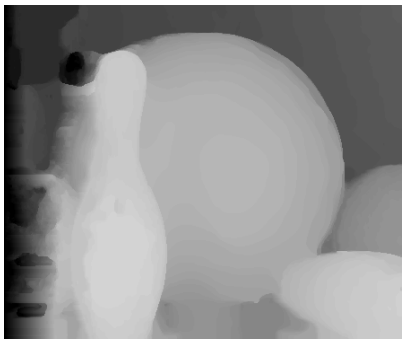
One goal of this thesis is providing a *unifying framework* for these semantic priors. To this end, we devise a generic primal-dual algorithm that can be used to minimize this class of energies.

Entropy Minimization for Continuous Multi-Label Optimization

Despite the enormous success of convex relaxation approaches in solving multi-labeling problems, these methods often suffer from the fact that computed solutions are far from being binary. Fractional (non-binary) solutions reflect the uncertainty of these algorithms and a posteriori rounding schemes often result in suboptimal results. In [9], we propose an



a) Semantic labeling on $\Omega \subset \mathbb{R}^2$



b) Geometric labeling on $\Omega \subset \mathbb{R}^2$

c) Semantic labeling on a 3D shape

d) Geometric labeling on $\Omega \subseteq \mathbb{R}^3 \times \mathbb{R}_+$

Figure 1.1: Examples of multi-labeling applied on different domain types. Image (a) depicts a semantic multi-label application on a highway image. Example (b) illustrates a depth-from-stereo reconstruction of a bowling scene where labels stand for different disparity values. The segmentation in (c) is performed on a 3D shape [158]. Image (d) depicts a frame from a spatio-temporal reconstruction [142] where the labels are object/void.

algorithm which uses the entropy of the objective variable as a measure of relaxation tightness, which in turn gives rise to more binary solutions than classical variational algorithms. The effectiveness of our approach is demonstrated on numerous computer vision applications, including multi-labeling, image inpainting, image segmentation, and spatio-temporal multi-view reconstruction.

In this thesis, we discuss augmenting a broader class of optimization problems with an entropy term, namely the class of *mixed integer programs*. This type of problems contains integer as well as continuous unknowns.

Joint Motion Estimation and Segmentation

In this cumulative thesis, we include a research paper [5] where we introduce a novel joint registration and segmentation approach for RGB-D images. Dynamic scenes are composed of rigid (e.g. , a tennis racket that is being swung by a tennis player) and non-rigid (e.g. , the tennis player's clothing that shifts and moves while he swings the racket) movements. The analysis of these movements has been a challenge to the field of motion estimation. Instead of assuming that a scene is composed of a number of independent rigidly-moving parts, our approach applies a smooth label transition in order to capture non-rigid deformations at transitions between the rigid parts of the scene. This is critical because our world is not composed solely of rigid movements.

1.1 Multi-Labeling

Multi-labeling is a highly ambiguous problem and the desired output heavily depends on its application. In this thesis, we mainly encounter multi-labeling in the context of image segmentation. However, applications including 3D and 4D reconstructions are also discussed. The goal of this section is to provide a mathematical setting for general multi-labeling independent of the type of the domain. Additionally, we present different application areas of multi-labeling demonstrating its versatility.

Formally we assume that the domain Ω is continuous and can be disjointly dissected as follows:

$$\Omega = \Omega_1 \cup \Omega_2 \cup \cdots \cup \Omega_n, \tag{1.1}$$

meaning that Ω constitutes of non-overalling parts $\Omega_i \subseteq \Omega$. The domain Ω depends on the application and can be:

- A 2D image, i.e. $\Omega \subset \mathbb{R}^2$
- A 3D volume, i.e. $\Omega \subseteq \mathbb{R}^3$
- A 4D volume, i.e. $\Omega \subseteq \mathbb{R}^3 \times \mathbb{R}_+$
- A 6D diffusion MRI volume, i.e. $\Omega \subseteq \mathbb{R}^3 \times \mathbb{R}^3$
- A non-Euclidean domain such as a shape manifold

Multi-labeling on different domain types is illustrated in Figure 1.1.

Mathematically, multi-labeling is the problem of seeking a label assignment $\mathcal{A}$ between the domain Ω and a discrete label space $\mathcal{L} = \{1, 2..., n\}$, i.e.

$$\mathcal{A} : \Omega \to \mathcal{L}. \tag{1.2}$$

This label assignment can be formulated as an integer programming problem on a continuous domain. This thesis will include applications where the label space encodes semantics, 3D reconstruction, spatio-temporal reconstructions and moving segments subject to different rigid body motions. The types of domain Ω we encounter in this thesis are usually

(a) Input image f (b) Image threshold u

Figure 1.2: Thresholding an image merely separates dark regions from bright ones. The application of such an approach is limited since typical objects of interest like the penguin consist of more than one color.

a subset of a Euclidean space, i.e. $\Omega \subset \mathbb{R}^N \times \mathbb{R}_+$. The major subject of this thesis is multi-labeling applied on images. In the next section, we give an overview of multi-label image segmentation algorithms by starting with the binary labeling case. We will assume $\Omega \subset \mathbb{R}^2$ for the rest of this chapter.

1.2 Image Segmentation via Energy Minimization

The most basic segmentation algorithm is taking a threshold of a input image $f : \Omega \to \mathbb{R}_+$ which yields the following binary function:

$$u(x) = \begin{cases} 1, & f(x) > \gamma, \\ 0, & \text{else}, \end{cases} \tag{1.3}$$

for some fixed constant $\gamma \in \mathbb{R}_+$. A thresholding result is illustrated in Figure 1.2. Although thresholding the image clearly separates dark and bright image areas, this simple approach fails to capture semantics such as separating the penguin from the background. Other heuristic methods for dissecting images are methods based on region merging [30], region growing [15] and the watershed transform [159]. While the aforementioned methods can be extremely fast, they lack a criterion of optimality. That is, without visual inspection, it is impossible for the computer to distinguish between a "good" and a "bad" solution. This chapter will address image segmentation methods based on energy minimization. These approaches often do not achieve global optimality. However, their energy gives us a mean to measure the quality of the results. Next, we will discuss the binary segmentation case and review local and global optimal approaches.

1.2.1 Contour based Methods

In this section, we elaborate on methods where the curve enclosing an object is explicitly formulated. One of the first energy based methods for image segmentation is the so-called snakes model introduced by Kass et al. in [93]. Given an input image $f : \Omega \to \mathbb{R}_+$, the curve evolution in the snakes approach can be described by the following energy:

$$E(C) = -\int_0^1 |\nabla f(C(s))|^2 \,\mathrm{ds} + \int_0^1 \left\{ \frac{\alpha}{2}|C_s(s)|^2 + \frac{\beta}{2}|C_{ss}(s)|^2 \right\} \mathrm{ds}, \tag{1.4}$$

where $C(s) : [0,1] \to \mathbb{R}^2$ denotes a parametric curve and α,$\beta \in \mathbb{R}_+$ are parameters. The last two terms make sure that the resulting curve is smooth, whereas the first term pushes the curve to the boundary of an object (i.e. ,large gradients). The snakes functional is minimized using a simple gradient descent scheme. The update steps are carried out on the curve by calculating a rate with which it transitions towards the inner normals. Classic active contour approaches depend inherently on the gradient information $|\nabla f|$ and the curve is thus prone to get stuck in edges which do not belong to an object. Additionally, the parametric representation of the curve does not allow topological changes such as splitting and merging. This is where the level set methods come into play.

1.2.2 Level Set Methods

Level set methods were introduced by Stanley Osher and J.A. Sethian in 1988 [141] and have been made popular in various areas, such as image processing, computer graphics, computational geometry and computational fluid dynamics. The most popular level set based segmentation algorithm is the Chan and Vese model [43] which relies on the idea of representing the curve C by the zero interface of a signed distance function $\varphi : \Omega \to \mathbb{R}$. The Chan and Vese functional for object/background segmentation can be formulated as follows:

$$\begin{aligned} E(\varphi) = \int_\Omega H(\varphi(x))f_o + (1 - H(\varphi(x)))f_b + |\nabla H(\varphi(x))| \,\mathrm{dx} \\ \text{s.t. } |\nabla\varphi(x)| = 1,\ \forall x \in \Omega \end{aligned} \tag{1.5}$$

where H denotes the Heaviside function

$$H(x) = \begin{cases} 0, & x < 0, \\ 1, & x \geq 0, \end{cases} \tag{1.6}$$

and f_o, f_b encode the costs of a pixel $x \in \Omega$ taking on label foreground ($H(\varphi(x)) = 1$) or background ($H(\varphi(x)) = 0$) respectively. Note that the constraint in (1.5) is known as the eikonal equation and needs to be fulfilled in order for φ to remain a signed distance function. This can be done by reinitializing φ after each iteration of the optimization algorithm [178]. The need to solve the boundary value problem in the constraint of (1.5) and the non-convexity of its objective make the approach suboptimal in terms of speed and global optimality of the solution.

1.2.3 Globally Optimal Methods for Binary Segmentation

In this section we discuss globally optimal methods for solving the task of binary segmentation. These formulations differ from contour based and level-set based methods in the following points:

- The problem representation no longer depends on an explicit formulation for the contour and therefore allows arbitrary changes in topology.
- Global optimality implying that the optimization of these cost functions is not dependent on the initialization.

There are mainly two lines of work solving the two-phase segmentation problem in a globally optimal way: models based on a discrete setting and methods developed under the assumption that the image domain is continuous.

Graph-Cut Methods

The image segmentation problem can be formulated in a discrete setting by interpreting the image domain as a discrete lattice. In the two-phase case, the resulting Markov random field (MRF) energy can be solved in various ways, including graph-cuts [24, 25, 76], linear programming [100] and message passing approaches [194]. The most efficient methods for solving binary MRF problems are graph-cut techniques. The only prerequisite is that their underlying energies are submodular. In that case a polynomial time algorithm can be applied in order to compute an s-t min-cut. Even more efficient is using the dual s-t max-flow formulation and solving it using a push-relabel strategy [69]. In addition to the binary labeling, strategies to generalize the graph-cut approach to multi-labeling have been explored. These move-making methods [26] solve iteratively binary graph-cut problems and refine the labeling until convergence. Being defined on a grid, discrete methods are inherently prone to exhibit metrification errors [95]. Additionally, they do not allow for sub-pixel accurate segmentation results. Finally, extensions to 3D and 4D domains are not trivial due to the extremely high memory consumption of these algorithms. Next, we present a method which is defined on a continuous domain and thus does not introduce metrification errors. Additionally, is it easily adaptable to higher dimensions and can be efficiently solved in parallel using GPUs.

Convex Relaxation Methods

Since the seminal work of Rudin, Osher and Fatemi [161] on image de-noising, variational methods have been popularized in the computer vision community. Especially the usage of the Total Variation (TV) has been proven very effective in reconstructing discontinuous signals. The first use of TV in image analysis dates back to the work of Shulman et al. [166] in the context of motion estimation. Interestingly TV has already been mentioned by the French researcher Camille Jorgen in 1881 [90] in a paper on Fourier series. The TV of a

function $u \in L^1(\Omega)$ is defined as follows [68]:

$$TV(u) := \sup_{\varphi} \left\{ \int_{\Omega} u \text{div } \varphi \, dx; \ \varphi \in C_0^1(\Omega, \mathbb{R}^2), \ ||\varphi||_\infty \leq 1 \right\}, \tag{1.7}$$

We define the space of functions of bounded variation as the space of all functions $u \in L^1(\Omega)$ which posses a bounded total variation, i.e.

$$BV(\Omega) = \left\{ u \in L^1(\Omega); \ TV(u) < \infty \right\}. \tag{1.8}$$

Total variation has been initially used as a tool for imposing spatial consistency on solutions. The geometric properties of TV where examined by Fleming and Federer [58] in their celebrated co-area formula which connects the total variation of a function u with the perimeter of all its level sets $u > \gamma$, $\forall \gamma \in \mathbb{R}$ as follows:

$$TV(u) = \int_{-\infty}^{\infty} Per(\{u > \gamma\}) d\gamma, \tag{1.9}$$

where $\{u > \gamma\}$ represents the level domain $\{x \in \Omega |\ u(x) > \gamma\}$ and where $Per(\{u > \gamma\})$ denotes the length of the interface of $\{u > \gamma\}$.

This insight from geometric measure theory has been a key idea to using TV in the context of image segmentation. If an object is represented by an indicator function u, TV(u) measures exactly its perimeter, i.e.

$$TV(u) = Per(\{u == 1\}), \tag{1.10}$$

where $\{u == 1\}$ denotes the set $\{x \in \Omega \,\big|\, u(x) = 1\}$. Using this insight, Chan, Esedoglu and Nikolova [138] devised a convex energy for solving the two-class segmentation problem which reads as follows:

$$E_{Bin}(u) = \lambda TV(u) + \langle \varrho, u \rangle_{L_2} \quad \text{s. t.} \quad u \in BV(\Omega; \{0, 1\}), \tag{1.11}$$

where $\varrho(x) = log \frac{p_f(x)}{p_b(x)}$ and where p_f and p_b are the probabilities of a point being foreground or background respectively. Parameter λ is a tuning parameter for penalizing the perimeter of the object represented by the indicator function u. Problem (1.11) is not convex due to the integer constraint $u \in BV(\Omega; \{0, 1\})$. Chan et al. [138] resolve the problem by a relaxation to the unit interval to obtain the following convex relaxed problem:

$$E_{Rel}(u) = \lambda TV(u) + \langle \varrho, u \rangle_{L_2} \quad \text{s. t.} \quad u \in BV(\Omega; [0, 1]). \tag{1.12}$$

In contrast to graph-cut methods, convex relaxation approaches do not introduce a grid bias and are trivially extendable to higher dimensions. Addionally, these methods are easily parallelizable.

Despite these advantages, formulation (1.12) has a catch, the integer constraints in the original binary problem (1.11) are not preserved and the solutions are possibly

fractional. However, using the so-called thresholding theorem Chan et al. [138] are able to prove that every thresholded solution

$$u_\gamma = \begin{cases} 1, & u^*(x) > \gamma, \\ 0, & \text{else}, \end{cases} \tag{1.13}$$

with:

$$u^* = \arg\min_u E_{Rel}(u), \tag{1.14}$$

is a solution of the original formulation (1.11) for any threshold $\gamma \in (0, 1)$. This also reveals the true nature of the original binary problem as it can exhibit more than one global solution.

1.3 Multi-Label Segmentation

Despite the global optimality of convex relaxation and graph-cut methods for the binary segmentation problem, they fail to capture the semantics of a scene with more than 2 labels. Recent segmentation algorithms tackle the case of having more than 2 labels. The multi-label segmentation problem is a combinatorial problem and is, in contrast to the 2 label case, NP-hard in general. The corresponding integer programming problem is naturally encoded by using indicator functions:

$$u_i(x) = \begin{cases} 1, & \text{if label } i \text{ is set in pixel } x \in \Omega, \\ 0, & \text{otherwise.} \end{cases} \tag{1.15}$$

By additionally imposing the uniqueness of labels in each point $x \in \Omega$, we obtain the following constraint:

$$\mathcal{S} := \left\{ u \in BV(\Omega, \{0,1\})^n \;\middle|\; \sum_{i=1}^{n} u_i(x) = 1 \quad \forall\, x \in \Omega \right\}. \tag{1.16}$$

Overall we obtain the generic energy

$$\min_{u \in \mathcal{S}} E(u) = E_D(u) + E_R(u) + E_S(u), \tag{1.17}$$

where $E_D(u)$ denotes the so-called dataterm which involves the input data and $E_R(u)$ the regularization term which ensures a spatially consistent solution. The term $E_S(u)$ can incorporate prior information, e.g. shape, geometry and co-occurrence statistics of labels. Energy (1.17) is very general and includes, in addition to spatially continuous methods, graph-cut based methods since their underlying linear programming (LP) problem can be cast in this very formulation. Energy (1.17) is non-convex due to the $\{0, 1\}$ constraint in (1.15) which makes it very hard to solve. Most algorithms resort to LP-relaxation techniques [119] where the integer constraint (1.15) is relaxed to

$$u_i(x) \in [0, 1] \quad \forall x \in \Omega, \;\; \forall i \in \mathcal{L}, \tag{1.18}$$

(a) Input image (b) Dataterm ϱ_1

Figure 1.3: The dataterm ϱ_1 for the foreground in a segmentation problem with $|\mathcal{L}| = 2$. This dataterm has been obtained using the method in [13].

and (1.16) transforms to the convex constraint:

$$\mathcal{S} := \left\{ u \in BV(\Omega, |0,1|)^n \;\middle|\; \sum_{i=1}^{n} u_i(x) = 1 \quad \forall\, x \in \Omega \right\}. \tag{1.19}$$

However, in contrast to the binary segmentation case, there exists no thresholding theorem. Hence, subsequent rounding of the solution does not yield optimal solutions to the original integer problem.

1.3.1 Dataterm

In the following, we discuss the dataterm E_D which depends on the input data. The dataterm assigns every point, x taking on a label i, a cost $\varrho_i(x)$. The overall cost is the integral over the domain Ω, i.e.

$$E_D = \sum_{i=1}^{n} \int_{\Omega} u_i(x) \varrho_i(x) \,\mathrm{dx}, \tag{1.20}$$

and can be calculated from color distributions [135], textural information [165] or depth cues [52]. More recent methods compute the dataterm using convolutional neural networks [209]. Note that in case $E = E_D$, the minimizer $u^* = \arg\min_u E(u)$ can be trivially computed as follows:

$$u_i^*(x) = \begin{cases} 1, & \text{if } i = \arg\min_i \varrho_i(x) \\ 0, & \text{otherwise.} \end{cases} \tag{1.21}$$

As can be seen in Figure 1.3, the dataterm is subject to noise and the accuracy of the labeling is not optimal. Therefore additional terms need to be introduced in order to impose spatial consistency.

1.3.2 Regularization

Many problems in computer vision are ill-posed, in the sense that often the solution is not unique, does not even exist or doesn't change continuously with the input or parameters. The typical remedy for ill-posed problems is to impose a regularization term which additionally endows the solution with a physically meaningful prior such as spatial consistency. In recent years, a multitude of works dealt with approximating the multi-label image segmentation problem. These include methods based on graph-cuts such as α expansion and α-β swap and approaches built upon the convex relaxation method described in Section 1.2.3. Continuous and discrete methods can both be cast into the formulation (1.17) and mainly differ in terms of the regularization term $E_R(u)$. Being defined on a continuous domain, convex relaxation methods have the advantage of not having a grid bias. Additionally, graph-cut approaches are based on refining the solution by solving a series of optimization problems sequentially whereas convex relaxation methods solve one convex optimization problem and can be easily optimized in parallel by state-of-the-art first-order approaches. The first to propose a convex relaxation for the multi-label problem in the continuous setting are Zach et al. in [206]. The authors use a label-wise TV regularization. Equivalent to formulation (1.7) the regularizer can be written as follows:

$$E_R(u) = \sup_{\varphi \in \mathcal{K}} \left\{ \sum_{i=1}^{n} \int_\Omega u_i \text{div } \varphi_i \, dx \right\}, \tag{1.22}$$

where $\mathcal{K}$ denotes the following convex constraint on the dual variable φ:

$$\mathcal{K} = \left\{ \varphi \in C_0^\infty(\Omega, \mathbb{R}^2)^n, ||\varphi_i||_\infty \leq \frac{1}{2}, \; 1 \leq i \leq n \right\}. \tag{1.23}$$

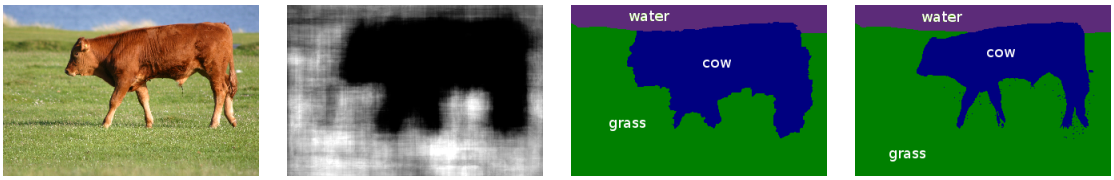
In [37] Chambolle, Cremers and Pock (CCP) introduce a convex formulation of the Potts model which is a tighter approximation than the energy introduced by Zach et al. The CCP formulation differs from the Zach formulation only in the constraint set $\mathcal{K}$. The CCP constraint set reads as follows:

$$\mathcal{K}_{CCP} = \left\{ \varphi \in C_0^\infty(\Omega, \mathbb{R}^2)^{n+1}, ||\varphi_i - \varphi_j||_\infty \leq 1, 1 \leq i < j \leq n+1 \right\}, \tag{1.24}$$

with $\varphi_{n+1} = 0$. Although exhibiting a tighter relaxation, the size of the CCP constraint grows quadratically with the number of labels, making it more difficult to perform a projection into it. In experiments, the tighter relaxation is hardly noticeable and occurs especially in pathological problems such as triple and 4-junctions of labels [37].

Non-Local TV

The above convex relaxations of the Potts model attempt to minimize the perimeter of each object in the scene. While this is a suitable prior for certain object classes, it can be suboptimal for structurally complex objects. For example objects with elongated structures cannot be recovered using TV due to its shrinking bias. Based on the seminal work of Gilboa et al. [67] on non-local differential operators for general graphs, Werlberger et



(a) Input image (b) Dataterm (c) TV regularization (d) Non local TV

Figure 1.4: Using non-local regularization (d) preserves elongated structures like the leg of the cow in contrast to the TV approach (c) which introduces a shrinking bias. The dataterm in (b) is associated with the label 'cow'.

al. [199] proposed using non-local TV in order to remedy the shrinking bias classical TV formulations exhibit. The key idea is to define label similarity of points x and y not only in a local scale but also considering points in a larger neighborhood. For this we assign each pair of points a weight w which measures their similarity:

$$w(x,y) = \exp\left[-\left(\frac{d_c(x,y)}{\alpha} + \frac{d_s(x,y)}{\beta}\right)\right]. \tag{1.25}$$

Here d_c and d_s encode some measures for color and spatial distances respectively. Parameters α and β are scaling parameters. The regularizer penalizes the weighted label differences between a point x compared to all its neighbors $y \in \Omega$:

$$E_R(u) = \sum_{i=1}^{n} \int_\Omega \sqrt{\int_\Omega w(x,y)\,(u_i(y) - u_i(x))^2\,dy}dx. \tag{1.26}$$

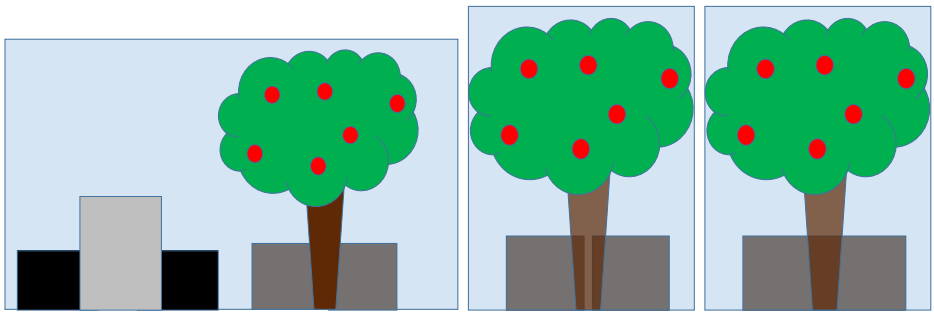
Figure 1.4 shows the difference between the classical Potts model and its non-local variant. One observes that the non-local smoothness penalty improves the boundary of the object by recovering elongated structures like the legs of the cow. In practice only a small neighborhood (e.g. a patch) $\mathcal{N}(x) \subset \Omega$ of pixel x is considered for calculating finite differences in (1.26).

1.3.3 Thresholding and Optimality Bounds

In order to recover a binary solution from the result of the convex relaxed optimization problem, one needs to resort to binarization techniques. One simple solution is to perform the following rounding scheme

$$\hat{u}_i(x) = \begin{cases} 1, & \text{if } i = \arg\max_i u_i^*(x) \\ 0, & \text{otherwise,} \end{cases} \tag{1.27}$$

where $\hat{u}$ denotes the rounded solution and u^* the solution of the relaxed problem. In contrast to the 2-label case, there exists no thresholding theorem. Hence, there is no a priori guarantee that the thresholded solution is optimal in terms of the original binary



(a) Observed Scene (b) 2 Boxes (c) 1 Box

Figure 1.5: Occam's principle suggests choosing the model where less boxes (1 box) are occluded by the tree over the more complex estimate of 2 boxes behind the tree.

energy. Chambolle et al. [37] proposed calculating an a posteriori bound based on the following relation:

$$E(\hat{u}) - E(\tilde{u}) \leq E(\hat{u}) - E(u^*), \tag{1.28}$$

where $\tilde{u}$ denotes the global minimizer of the original integer programming problem. The relation (1.28) is based on the fact that $E(u^*) \leq E(\tilde{u}) \leq E(\hat{u})$. Since the unknown energy gap $E(\hat{u}) - E(\tilde{u})$ of the rounded solution $\hat{u}$ to the true binary minimizer $\tilde{u}$ is bounded by the known gap $E(\hat{u}) - E(u^*)$ we obtain the optimality measure:

$$E(\hat{u}) - E(u^*). \tag{1.29}$$

Hence, a smaller bound is an indication of the proximity to the minimizer of the binary problem.

Similar to the a priori bound provided for α-expansion in [26], Lellmann et al. [109] proposed a probabilistic rounding scheme which provides an a priori optimality bound for continuous multi-labeling. In practice, this probabilistic rounding method happens to be slow and does not provide visually satisfactory solutions. In [9], we proposed minimizing the entropy of the unknowns in order to penalize deviations from the integer constraint (1.15). Although our approach is not convex, our algorithm is guaranteed to converge and experimental results show that we consistently obtain better optimality gaps compared to pure convex relaxation approaches. In Chapter 7, we present a generalizing framework in which we introduce entropy minimization to the broader class of mixed-integer programs.

1.4 Incorporating Prior Information

The problem of dissecting an image into regions is a highly ambiguous task and often one needs to provide the algorithm with additional information in order to achieve better results. In the following, we discuss two different types of higher level priors, namely geometric and label configuration or semantic priors.

1.4.1 Geometric Priors

The most basic geometric prior forces the solution to be spatially smooth. Spatial smoothness in the in context of image segmentation reflects the prior expectation of objects to have minimal boundary. Boundary length priors arrised initially in the work of Mumford and Shah [131] and have been adopted in various forms, including explicit [93] and implicit [138] representations of objects. A more sophisticated way of regularity is devised in [80] and can be considered as an object class specific smoothing term which implements a surface orientation prior.

Another line of work on geometric priors has promoted global priors involving all points in Ω. One of the first global geometric priors is the balooning force introduced by Cohen et al. [46] which is designed to avoid local minimizers in the active contour model. In [96, 97], Klodt et al. introduced a family of shape priors to binary segmentation in the form of moment constraints. These convex constraints range from simple area constraints to complex higher order moments. By increasing the order, the authors are able to encode simple shape priors. In [136] Niewenhuis et al. propose coupling the volume of different labels in order to enforce learned segment proportions. In [116, 172], Strekalovskiy et al. proposed different penalization of label jumps for different directions. These penalties are capable of encoding complex label layout priors. Using dynamic programming, a similar approach has been devised in the discrete setting [60].

A more direct way of realizing shape priors is to encourage the consistency between a learned silhouette and the segmentation result [44, 47]. However, these methods suffer from local optimality and a good initialization is necessary to obtain optimal results. A remarkable exception is the work of Schoeneman et al. [163] where a polynomial time algorithm for matching shapes to images is proposed.

1.4.2 Label Configuration Priors

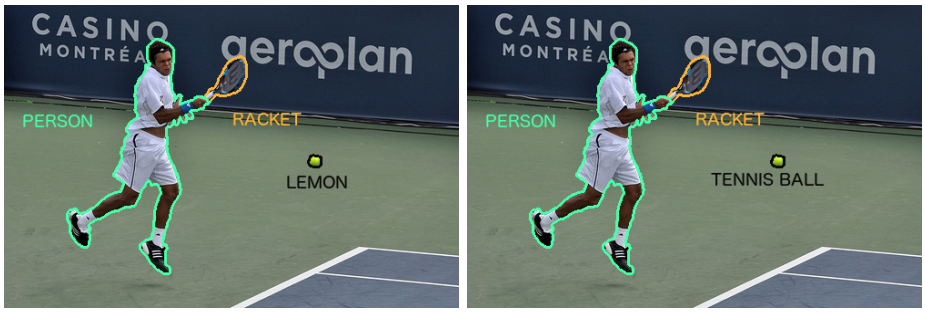
Label configuration priors differ from geometric priors in the following criteria:

- Label configuration priors are independent of location and size of the regions.
- They depend on the occurrence of certain label configurations.

In the following we discuss some instances of label configuration priors, namely the minimal description length, the co-occurrence and a hierarchy-based prior.

Minimal Description Length

Minimal description length (MDL) is the most basic label configuration priors and and derives from the criterion of Occam's razor [123]. Occam's razor states that simpler explanations are preferred over more complex ones. It has been stated by earlier philosopher John Duns Scotus in 1639 "Pluralitas non est ponenda sine necessitate" ("Plurality is not to be posited without necessity") in Opera Omnia. This principle of parsimony can also be applied in data sciences by choosing the simplest model which fits the data in order



(a) (b)

Figure 1.6: The co-occurrence prior helps correcting label configurations. After wrongly labeled as a lemon (a) the tennis ball is recovered in (b) and the label configuration "person, racket, tennis ball" is favoured.

to avoid overfitting. Figure 1.5 illustrates Occam's razor where due to occlusions, it is not clear how many boxes are in the scene. The answers range from 3 boxes to infinitely many boxes. Occam's razor encourages making less hypothesizes which conforms to human perception.

The same principle has been applied to the task of image segmentation [50, 106, 204, 211]. In addition to the data and smoothness term, a penalty for favoring a minimal description length representation of the scene is introduced. In [11], we develop a generalized version of the MDL prior by extending its classical linear label count penalty to a composition with convex monotone functions. This way, we can e.g. impose an upper bound on the number of emerging labels. The MDL principle is often suboptimal for obtaining good segmentation results. This is due to the fact that it merely penalizes the label count and is agnostic to the emerging label combinations.

Co-Occurrence Prior

A much more fine-grained approach is to penalize configurations of labels which are unlikely to occur by incorporating learned label statistics. An illustration of the co-occurrence prior can be seen in Figure 1.6. This prior helps correcting the labeling by encouraging labels 'person', 'racket' and 'tennis' to co-occur and by penalizing the emergence of labels 'person', 'racket' and 'lemon'. This prior has been already incorporated by Rabinovich et al. [154] where after segmentation and object categorization, learned co-occurrence information is used in a final refining step. Moreover, Ladicky et al. [104] incorporate co-occurrence potentials into a conditional random field. The resulting energy is then minimized by means of α expansion and $\alpha - \beta$ swap [26]. In [10], we integrate the idea of incorporating the co-occurrence penalty into continuous multi-label optimization.

(a) Label hierarchy (b) Office scene (c) Desert scene

Figure 1.7: Different scenes influence the type of labeling occurring in the segmentation. Depending on contextual information provided as a label hierarchy (a) the algorithm chooses from office labels (b) or assigns labels from the desert context (c).

Hiarchical Prior

While the co-occurrence prior is very powerful, it still depends on pre-learning from a database. A more principled prior which is already integrated in human vision is the ability to group labels under a hierarchy. To this end, a hierarchy of objects is constructed by introducing superordinate scene labels. This context specific prior has been introduced by Torralba et al. [187] where the recognized scene type penalizes non-conforming configurations. A more holistic approach is presented by Delong et al. [50] in the discrete setting which yields a sequential algorithm. In [8], we introduce a hierarchical prior into variational multi-labeling. We are able to devise a single optimization problem incorporating a hierarchy constructed a priori. Figure 1.7 illustrates an example hierarchy with two different scenes. In an ideal case an algorithm should be able to recognize the scene and restrict the label set to the labels conforming with it.

In Chapter 6, we introduce a unifying framework which generalizes our works [8, 10, 11] on semantic priors for multi-label segmentation. To this end we present a generic first order primal-dual algorithm for semantic segmentation.

Chapter 2

Outline of the Thesis

The present cumulative thesis is divided into three parts. In the following chapter we provide a summary for each part.

Part I:

Introduction

The first chapter gives a motivation and an introduction to the general concept of multi-labeling. Next, we provide an overview of classical image segmentation methods based on energy minimization and state-of-the-art multi-label algorithms. Finally, we elaborate on the different priors commonly used in image segmentation algorithms including priors of geometric and semantic nature.

Discretization

Variational methods are initially formulated on a continuous domain. In order to implement these methods on a digital computer, one needs to discretize the image domain and the involved differential operators. In this chapter we provide example discretizations of typical differential operators by means of a matrix vector representation.

Mathematical Optimization

Since we are mainly dealing with algorithms based on numerical optimization, we introduce in this part basic concepts from convex optimization such as convexity and duality. Additionally, we elaborate on a well known iterative optimization method, namely the proximal-point algorithm. As an instance of the proximal-point algorithm we derive one of the most popular first order optimization algorithm, namely the primal-dual algorithm [39, 147] which will be used throughout this thesis.

Unifying Framework for Semantic Multi-Labeling

This chapter aims at providing a unifying framework for the algorithms we presented in [10] and [8] which are included in Chapters **??** and 9 respectively. These methods altogether deal with semantic priors i.e. co-occurrence and hierarchical priors. Conceptually these algorithms are very similar which inspired us to provide a generalizing algorithmic framework. Starting with a generic semantic multi-labeling problem, we derive a primal-dual algorithm which can handle a family of multi-labeling problems with semantic priors. We apply our algorithm to multi-label segmentation with a minimal description length prior which imposes an upper bound on the label count.

Entropy Minimization for Mixed Integer Programs

In [9], we introduce adding Shanon's entropy to convex relaxation methods. Our approach delivers solutions which are more binary and which have a tighter energy than pure convex relaxation methods. In this chapter we expand our approach to the broader class of mixed integer programs. Mixed integer programs contain integer as well as continuous unknowns. In the same spirit of [9], we augment the convex relaxation of the mixed integer problem with an entropy term resulting in a convex concave problem. Using tools from DC-Programming, we are able to devise a convergent algorithm for solving this problem. As an example application we apply our approach on the problem of image cartooning where we minimize the L_0 norm of the image gradient along with a data fidelity term. Experimental results demonstrate that solving this optimization problem leads to a piece-wise constant approximation of the input image. Comparative results show that our algorithm competes with state-of-the-art cartooning algorithms.

Part II:

This part includes a selection of peer reviewed research papers that were published during this thesis including:

- Co-occurrence Priors for Continuous Multi-labeling [10]
- Convex Optimization for Scene Understanding [8]
- Entropy Minimization for Convex Relaxation Approaches [9]
- Smooth Piece-Wise Rigid Scene Flow from RGB-D Images [5]

We additionally provide a short summary for each included publication.

Part III

Concludes the thesis and proposes directions for possible future works.

Chapter 3

Main Contributions

The general aim of this thesis is to introduce different convex priors to continuous multi-label segmentation. The resulting optimization problems remain convex and can be solved using state of the art first order methods. Additionally, we introduce a mean to solve mixed integer programs by adding an entropy term to the convex relaxed problem and thus improving its integrality. Finally, a framework for joint motion estimation and segmentation is presented. Overall, the contributions of this thesis can be summarized as follows:

Introducing Convex Semantic Priors to Multi-Labeling

The major contribution of this thesis is introducing label configuration priors to continuous multi-labeling. The simplest label configuration prior is the so called minimum description length prior. By applying arbitrary convex and monotonously increasing functions on the label count, we generalize the classical linear penalization introduced in the discrete [51] and the continuous [203] setting. This allows e.g. imposing an upper bound on the label count. Furthermore, we introduce a co-occurrence prior within a spatially continuous framework. In contrast to sequential discrete methods [104], our formulation [10] does not have a grid bias and can be solved optimally in parallel within one optimization problem. A more principled approach to introducing semantic prior information is imposing a natural hierarchy on the labels. This allows inferring superordinate scene labels alongside object labels. In contrast to solving a sequence of discrete problems via fusion moves as in [50], we devise an algorithm [8] which imposes a hierarchy prior by solving one convex optimization problem which does not introduce metrification errors.

In Chapter 6, we provide a *unifying framework* for all our published works on semantic multi-labeling.

Entropy Minimization for Multi-Label Optimization

We propose introducing the entropy of the objective variable in order to measure the tightness of the relaxed solution in convex relaxation methods. By jointly optimizing the

entropy and the original convex relaxation we are able to improve on the integrality of the solutions. The resulting convex-concave procedure can be solved using a specialized DC-Programming algorithm which includes a state of the art primal-dual subroutine. In various experiments and in theory, we are able to show that the obtained solutions exhibit improved integrality and tighter energy bounds compared to state of the art convex relaxation methods. The proposed method does not add up to the complexity since, in practice, only a few iterations of the primal-dual subroutine are sufficient in order to obtain satisfactory results. Considered applications include multi-label image inpainting, image segmentation and spatio-temporal multi-view reconstruction. This approach has been published in a research paper [9] which is included in Chapter 10.

In Chapter 7, we generalize this approach by introducing an entropy term to a convex relaxation of the more general class of mixed-integer programs.

Joint Motion Segmentation and Estimation

In our published research paper [5], we tackle the problem of jointly estimating motion and segmenting the scene in RGB-D images. Motion estimation as well as image segmentation are one of the most fundamental problems in computer vision. Although the problem formulations are different, both tasks are highly interdependent since moving objects are usually spatially consistent and can be dissected by image segmentation algorithms. This is were our motion cooperation (MC-Flow) algorithm comes into play, that is, our formulation is able to jointly segment the scene and infere for each segment its underlying rigid body motion in 3D. Our approach is the first to perform soft labelling and thereby allow an interpolation between rigid body motions. Thus, we are able to even recover non-rigidly moving parts in the scene which allows for more accurate motion estimations. We are able to outperform state of the art scene flow algorithms in terms of quantity and quality. This research paper is included in Chapter 11.

Research Papers not included in this Thesis

In addition to the works included in Part II, a variety of peer reviewed papers not included in this thesis have been published. These include works on:

1. Medical imaging [2, 3], where the goal is to reconstruct data in 6-dimensional image×diffusion space in diffusion magnetic resonance imaging.
2. Real time RGB-D scene flow estimation [4], where we devised a method for computing dense scene flow in real time using a first order primal dual algorithm which we implemented on a GPU.
3. Unsupervised image segmentation [6], where we perform the separation of figure and ground based on their mutual information. To this end, a strategy of minimizing a sequence of convex problems is applied in order to solve the arising non-convex optimization problem.

A full list of our research papers can be found within the references.

Chapter 4

Mathematical Optimization

The energy formulations mentioned in Section 1.3 are convex relaxations of NP hard problems and are therefore amenable to the powerful tools of convex analysis. Convex analysis emerged in the 1960s from works by R. Tyrrel Rockafellar [157], Jean-Jacques Moreau [130] and Werner Fenchel [61] who extended studies on convex sets to convex functions and how to minimize them. It is a field dealing with convex functions and the characterization of their minima. Assessing the "hardness" of optimization problems becomes just a matter of wether it is convex or not. Additionally, the important concept of duality is a mean to gain a different perspective on the problem and provide a lower bound to its energy. The canonical problem in convex optimization can be stated as a convex objective function F subject to a convex constraint $\mathcal{C}$, i.e.

$$\begin{aligned} &\text{minimize } F(u) \\ &\text{subject to } u \in \mathcal{C}, \end{aligned} \tag{4.1}$$

where we assume $F : X \to \mathbb{R}$, $\mathcal{C} \subset X$ and $u \in X$ with X being a finite dimensional vector space with $n = \dim X$. It is useful to allow infinite values for the range of functions in order to examine their limit behavior and to allow incorporating constraints into the objective. For this we extend the real line by ∞ and define the mapping:

$$E : X \to \mathbb{R} \cup \{\infty\}. \tag{4.2}$$

By extending F by the indicator function of $\mathcal{C}$ we obtain

$$E(u) = F(u) + \delta(u \in \mathcal{C}), \tag{4.3}$$

with

$$\delta(u \in \mathcal{C}) = \begin{cases} 0, & u \in \mathcal{C}, \\ \infty, & \text{else.} \end{cases} \tag{4.4}$$

This means that the objective E attains an infinite value if the constraint is not feasible. In order to recover the original domain and feasible solutions we introduce the notion of effective domain:

$$\operatorname{dom} F = \{u : E(u) < \infty\}. \tag{4.5}$$

4.1 Convexity

A convex function E satisfies the following condition:

$$E(tu + (1-t)v) \leq tE(x) + (1-t)E(v), \tag{4.6}$$

for any $t \in [0,1]$ and $u, v \in X$. The convexity of a function means that every critical point $\bar{u}$ is a global minimum. In general, a convex function does not necessarily have a unique minimum. One way to show a function E admits a single minimizer is strict convexity, hence if the inequality in (4.6) is strict. The uniqueness of the solution does not imply strict convexity, e.g. the function $E(u) = |u|$ is only minimized at 0 although not strictly convex.

4.2 Existence of the Solution

The convexity of a function is a necessary condition to find a global minimum. However, a convex function does not necessarily have a minimum. A convex function E admits a minimizer if it exhibits the following properties:

- *Lower semi-continuity*:

  For every sequence $u_k \to u$ we have:

  $$E(u) \leq \liminf_k E(u_k). \tag{4.7}$$

  The geometric interpretation of lower semi-continuity is the closedness of the epigraph of E.

- *Coercivity*:

  A function f is called coercive if it satisfies the following condition:

  $$\lim_{||u|| \to \infty} E(u) = \infty. \tag{4.8}$$

4.3 Optimality Conditions

The optimality conditions of an optimization problem serve as a tool for identifying a solution and computing a descent direction. Since we deal with energy functions which are not always differentiable, the optimality condition for a global minimum needs to be written with respect to the subdifferential. The necessary condition for u^* to be the minimizer of E reads as follows:

$$u^* = \arg\min_u E(u) \Leftrightarrow 0 \in \partial E(u^*), \tag{4.9}$$

where $\partial E(u^*)$ denotes the subdifferential of E at point u^*. The subdifferential is a set valued operator containing all subgradients, i.e.

$$\partial E(u) = \{p \in X^* | E(v) - E(u) - \langle p, v - u \ \geq 0\rangle, \forall v \in X\}, \tag{4.10}$$

where X^* denotes the dual space of X. Equation (4.9) will serve in Section 4.7 as a starting point for deriving a first-order optimization method from the conceptional proximal-point algorithm which computes the zeros of general monotone operator.

4.4 Duality and the Conjugate Function

The principle of duality and the Fenchel conjugate are the most important tools in convex analysis. The counterpart to the Fourier transform in convex analysis is the Fenchel conjugate function E^* which can be written as follows:

$$E^*(p) = \sup_{u \in X} \langle p, u\rangle - E(u). \tag{4.11}$$

The conjugate function is also known as the Legendre transform and measures the maximal distance between the linear form $\langle u, p\rangle$ and $E(u)$. In the 1D case, the conjugate function is illustrated in Figure 4.1 where p corresponds to the slope of $E(u)$ at point u and where $E^*(p)$ is its intersection with the v axis. Similar to the Fourier transform, the conjugate function can be used to facilitate the computation of the convolution (inf-convolution) in the min-plus algebra [144]. The Fenchel conjugate often appears while reformulating the Lagrangian of an optimization problem. This makes it easier to identify certain functions and simplify optimization problems. An important feature of conjugate functions is that non-convex problems can be rendered convex by computing the double conjugate function E^{**}, i.e.

$$E^{**}(u) = \sup_{p \in X} \langle p, u\rangle - E^*(p). \tag{4.12}$$

Interestingly, $E(u) = E^{**}(u)$ if E is convex and lower semi-continuous. Furthermore, E and E^{**} always share the same set of minimizers in case E admits any. Geometrically the epigraph of E^{**} is the convex hull of the epigraph of E.

4.5 Canonical Variational Problem

Most variational problems in computer vision can be stated in a structured way in the sense that they can written as the separable function

$$E(u) = F(Ku) + G(u), \tag{4.13}$$

where $K : X \to Y$ is a continuous linear operator mapping from X to a finite-dimensional vector space Y with $\dim Y = m$, and where F and G are proper convex lower semi-continuous functions. Energy (4.13) can be reformulated as the constrained formulation

$$E(u) = F(v) + G(u) \qquad \text{s. t.} \quad v = Ku. \tag{4.14}$$

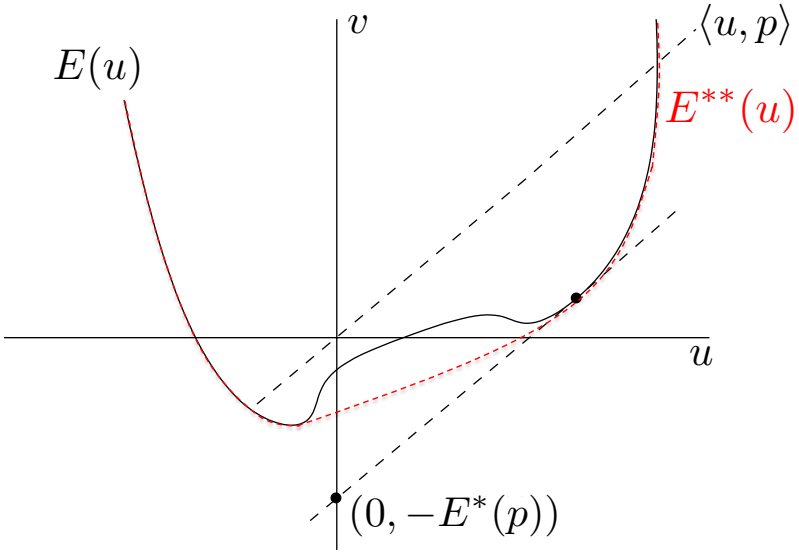


Figure 4.1: The conjugate function evaluated at point p is the maximum gap between the linear functional $\langle u, p \rangle$ and $E(u)$. For a differentiable function this case occurs exactly at $(u, E(u))$ where $\frac{\partial E(u)}{\partial u} = p$. The epigraph of the double conjugate function E^{**} (dashed red line) is the convex envelope of the epigraph of E.

The associated Lagrangian to (4.14) reads as:

$$L(u, v, p) = \inf_{u,v} F(v) + G(u) + \langle p, Ku - v \rangle. \tag{4.15}$$

The so called dual problem $D(u)$ is only dependent on multiplier p and can be derived from the Lagrangian problem by taking its infimum with respect to the primal variables, i.e.

$$D(p) = \inf_{u,v} L(u, v, p). \tag{4.16}$$

Problem (4.16) can be written explicitly by performing the following reformulations:

$$\begin{aligned} D(p) &= \inf_{u,v} F(v) + G(u) + \langle p, Ku - v \rangle \\ &= \inf_{u} \{G(u) + \langle p, Ku \rangle\} + \inf_{v} \{F(v) - \langle p, v \rangle\} \\ &= -\sup_{u} \{-G(u) - \langle p, Ku \rangle\} - \sup_{v} \{-F(v) + \langle p, v \rangle\} \\ &= -\sup_{u} \left\{-G(u) - \langle K^T p, u \rangle\right\} - \sup_{v} \{-F(v) + \langle p, v \rangle\} \qquad (4.17) \\ &= -G^*(-K^T p) - F^*(p). \qquad (4.18) \end{aligned}$$

The resulting dual problem has a closed form. Note that in (4.17) we identified the Fenchel conjugates G^* and F^*. The dual formulation is very useful for understanding the structure of the primal problem. Additionally, the dual problem is often easier to solve than its primal counterpart.

State-of-the-art approaches in numerical optimization tackle the so-called saddle-point problem. The reason is, that in the so called primal-dual formulation the optimization process can be accelerated by exploiting information of both the primal and the dual variable. The primal-dual formulation can be derived by substituting F by its double conjugate F^{**} in (4.13) and the resulting saddle-point problem takes the form:

$$\min_u \max_p \quad G(u) - F^*(u) + \langle p, Ku \rangle. \tag{4.19}$$

4.6 Primal-Dual Gap

The dual energy $D(p)$ is a lower bound on the primal energy, i.e.

$$D(p) \leq E(u) \quad \forall p \in Y, u \in X. \tag{4.20}$$

This even applies for non-convex functions. Furthermore, for an optimal primal-dual pair ($u^* = \arg\min E(u), p^* = \arg\max D(p)$) both the dual and the primal energies coincide (i.e. $E(u^*) = D(p^*)$) under the condition of strong duality [23]. Under the same condition, if a primal-dual optimal pair is achieved the so-called primal-dual gap

$$\mathcal{G}(u, p) = F(Ku) + G(u) + G^*(-K^T p) + F^*(p), \tag{4.21}$$

vanishes. The primal-dual gap provides a natural convergence criterium for solving the saddle-point formulation (4.19) and is much more expressive than measuring the decrease of energy during an iterative algorithm.

4.7 Proximal-Point Algorithm

Most first-order methods can be shown to be a special case of the proximal-point algorithm (PPA) [156]. This is of great importance since proving the convergence of these methods follows from the convergence of the PPA. The PPA is a fixed point method which serves as a mean of finding zeros of a maximal monotone operator $\mathcal{T}$, i.e. finding a u^* such that:

$$0 \in \mathcal{T}(u^*). \tag{4.22}$$

An operator is a general set valued mapping:

$$\mathcal{T} : \text{dom}\, \mathcal{T} \to 2^X, \tag{4.23}$$

where 2^X denotes the set of all subsets of X and dom $\mathcal{T} = \{x | \exists y \, (x, y) \in \mathcal{T}\}$. More specifically, we can define an operator as the following relation:

$$\mathcal{T} = \{(x, y) \in X \times X | \, y \in \mathcal{T}(x)\}, \tag{4.24}$$

and we call an operator monotone if it fulfills the following condition:

$$\langle x - v, y - w \rangle \geq 0 \qquad \forall (x, y), (v, w) \in \mathcal{T}. \tag{4.25}$$

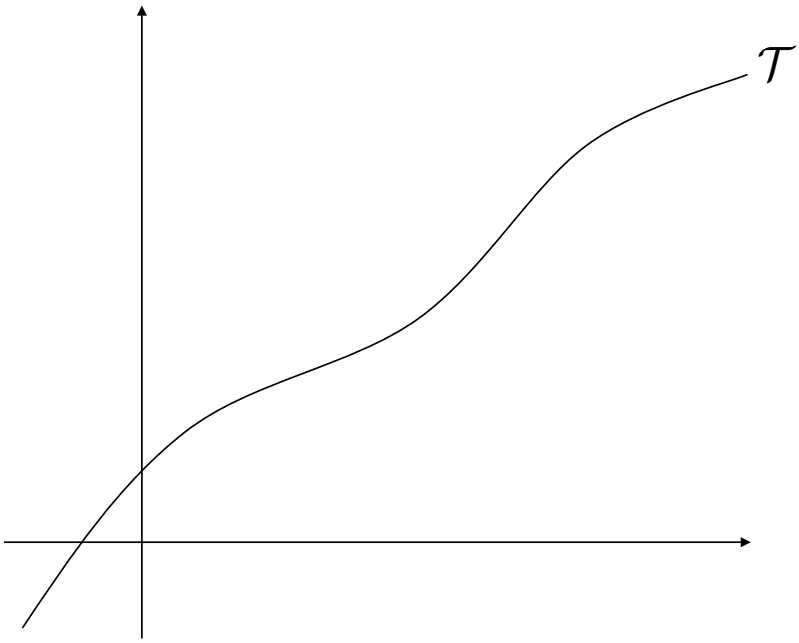


Figure 4.2: An example of a maximal monotone operator. In 1D an operator is maximal monotone if it is a curve with non-negative slope

A maximal monotone operator $\mathcal{T}$ is a monotone operator such that, there is no other monotone operator $\tilde{\mathcal{T}}$ with $\mathcal{T} \subset \tilde{\mathcal{T}}$. An example of a monotone operator where $X = \mathbb{R}$ is depicted in Figure 4.2. Geometrically speaking, a maximal monotone operator is a curve with positive slope.

The proximal-point algorithm generates the following implicit sequence in order to solve (4.22):

$$0 \in \mathcal{T}(u^{n+1}) + \mathcal{M}(u^{n+1} - u^n). \tag{4.26}$$

The proximal algorithm can be written with respect to any weighted norm $|\cdot|_{\mathcal{M}}$. However, the convergence of the algorithm is only guaranteed if matrix $\mathcal{M}$ is a symmetric and positive definite matrix. As a consequence, the optimization can be considered to be carried out in a Hilbert space endowed with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{M}}$. Rockafellar et al. prove in [156] the convergence of the proximal-point algorithm for general Hilbert spaces which applies to this case. We can make the above equation explicit by solving for u^{n+1} and we obtain the following iterative fix-point procedure:

$$u^{n+1} = (I + \mathcal{M}^{-1}\mathcal{T})^{-1}(u^n). \tag{4.27}$$

The operator $(I + \mathcal{M}^{-1}\mathcal{T})^{-1}$ is single valued [127] and is known as the resolvent of $\mathcal{T}$. The iterates of the proximal-point algorithm for the 1D case are depicted in Figure 4.3 where the scalar $\mathcal{M}^{-1} = \tau$ corresponds to the step size of the algorithm. Computing the resolvent operator is often as hard as solving $0 \in \mathcal{T}$. In order to circumvent this, Rockafellar et al. proposed solving the resolvent operator approximately [156]. However,

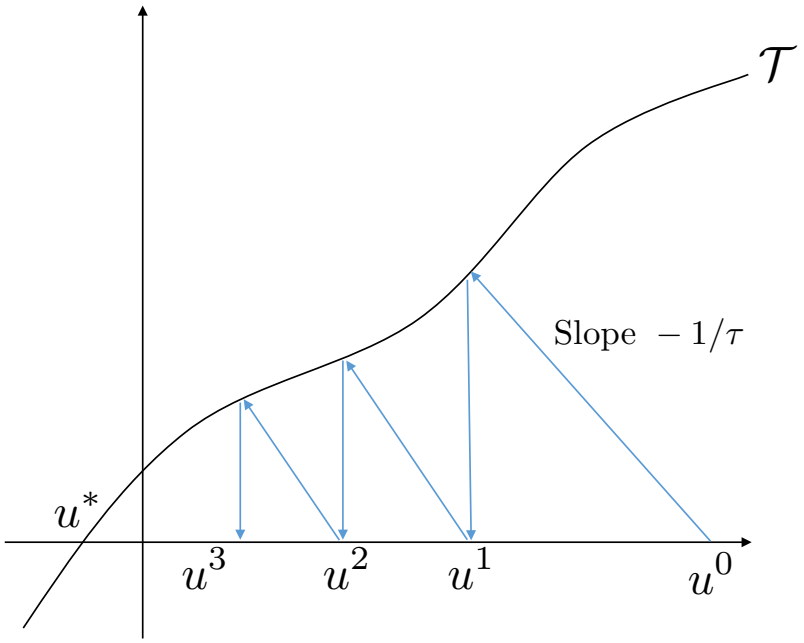


Figure 4.3: The iterates of the proximal-point operator for a monotone operator [54]. It is obvious that bigger step sizes lead to a faster convergence since the slope of the iterates decreases leading to iterates pointing more to the zero of $\mathcal{T}$.

the convergence of such approximate algorithm is only guaranteed if the error is summable. The most common strategy to solving (4.22) is to examine whether $\mathcal{T}$ can be split as a sum of two operators:

$$\mathcal{T} = A + B. \tag{4.28}$$

Fortunately, most variational problems in computer vision can be cast in this form since they typically contain a sum of terms such as the dataterm and the regularization.

4.8 Primal-Dual Method for Convex Problems

In the following we derive a first-order primal-dual algorithm proposed in [39, 147] for solving optimization problems in the form (4.13). This algorithm is extensively used in the works introduced in this thesis. This is due to the fact that first-order methods are suitable for the large scale optimization problems we encounter. Additionally, they are easily parallelizable which facilitates an optimized GPU implementation. For a first-order method, the primal-dual algorithm exhibits optimal convergence behavior for the class of non-smooth problems.

One important example of monotone operators is the set valued subdifferential mapping $\partial E(u)$ which is maximal monotone if and only if E is convex and lower semi-continuous. If we consider the problem in (4.19) and calculate its subdifferential the optimality condition

for characterizing a saddle points $(\hat{u}, \hat{p}) \in X \times Y$ can be written as follows:

$$\begin{aligned} 0 &\in K^T \hat{p} + \partial G(\hat{u}) \\ 0 &\in K\hat{u} - \partial F^*(\hat{p}). \end{aligned} \tag{4.29}$$

Hence, in order to cast our optimization problem in (4.26) we identify $\mathcal{T}$ with

$$\mathcal{T}(p, u) = \begin{cases} K^T p + \partial G(u) \\ Ku - \partial F^*(p) \end{cases}. \tag{4.30}$$

Next we split the operator $\mathcal{T} = A + B$ as in (4.28) by setting:

$$A = \begin{pmatrix} \partial F^* & 0 \\ 0 & \partial G \end{pmatrix}, \quad B = \begin{pmatrix} 0 & K^T \\ -K & 0 \end{pmatrix}, \tag{4.31}$$

we obtain the following iterative procedure for the proximal-point algorithm

$$0 \in (A + B) \begin{pmatrix} u^{k+1} \\ p^{k+1} \end{pmatrix} + \mathcal{M} \begin{pmatrix} u^{n+1} - u^n \\ p^{n+1} - p^n \end{pmatrix}. \tag{4.32}$$

Furthermore we set the preconditioning Matrix $\mathcal{M}$ to be as follows:

$$\mathcal{M} = \begin{pmatrix} T & -K^T \\ -K & \Sigma \end{pmatrix}. \tag{4.33}$$

For the sake of simplicity, we further split $\mathcal{M}$ in $\mathcal{M} = N + M$ with

$$M = \begin{pmatrix} T & 0 \\ 0 & \Sigma \end{pmatrix}, \quad N = \begin{pmatrix} 0 & -K^T \\ -K & 0 \end{pmatrix}. \tag{4.34}$$

After substituting $\mathcal{M}$ in (4.32) we obtain the following implicit equation:

$$0 \in (A + B) \begin{pmatrix} p^{k+1} \\ u^{k+1} \end{pmatrix} + (M + N) \begin{pmatrix} u^{k+1} - u^k \\ p^{k+1} - p^k \end{pmatrix} \tag{4.35}$$

$$(A + M) \begin{pmatrix} u^{k+1} \\ p^{k+1} \end{pmatrix} \in M \begin{pmatrix} u^k \\ p^k \end{pmatrix} - B \begin{pmatrix} p^{k+1} \\ u^{k+1} \end{pmatrix} - N \begin{pmatrix} p^{k+1} - p^k \\ u^{k+1} - u^k \end{pmatrix}. \tag{4.36}$$

By solving for (p^{k+1}, u^{k+1}) we transform above implicit equation into explicit iterates

$$\begin{pmatrix} u^{k+1} \\ p^{k+1} \end{pmatrix} \in (A + M)^{-1} \left(M \begin{pmatrix} u^k \\ p^k \end{pmatrix} - B \begin{pmatrix} u^{k+1} \\ p^{k+1} \end{pmatrix} - N \begin{pmatrix} u^{k+1} - u^k \\ p^{k+1} - p^k \end{pmatrix} \right). \tag{4.37}$$

If we substitute B and N with their values in (4.34) we obtain

$$\begin{pmatrix} u^{k+1} \\ p^{k+1} \end{pmatrix} \in (A + M)^{-1} \left(M \begin{pmatrix} u^k \\ p^k \end{pmatrix} + \begin{pmatrix} K^T p^{k+1} + K^T p^{k+1} + K^T p^k \\ Ku^{k+1} + Ku^{k+1} - Ku^k \end{pmatrix} \right). \tag{4.38}$$

By further simplifying and substituting for A and M we obtain

$$\begin{aligned} u^{k+1} &= (I + T^{-1}\partial G)^{-1}\left(u^k - T^{-1}K^T p^k\right) \\ p^{k+1} &= (I + \Sigma^{-1}\partial F^*)^{-1}\left(p^n + \Sigma^{-1}K(2u^{k+1} - u^k)\right). \end{aligned} \tag{4.39}$$

The resolvent operators $(I + T^{-1}\partial G)^{-1}$ and $(I + \Sigma^{-1}\partial F^*)^{-1}$ are called the proximity operators of G and F^* respectively and are single valued. Hence one can replace the inclusion $\in$ with an equality going from (4.38) to (4.39). Interestingly the resolvents of monotone differential operators can be calculated by solving an optimization problem. For example the proximity operator of G can be rewritten as follows:

$$(I + T^{-1}\partial G)^{-1}(\hat{u}) = \arg\min_u G(u) + \frac{1}{2}\left(\langle u - \hat{u}, T(u - \hat{u})\rangle\right)^{\frac{1}{2}}, \tag{4.40}$$

where $(\langle u - \hat{u}, T(u - \hat{u})\rangle)^{\frac{1}{2}}$ denotes the weighted distance of u and $\hat{u}$ by the preconditioning matrix T. By updating u^{k+1} and p^{k+1} alternatively one arrives at an implicit update scheme. The overall primal-dual algorithm [39, 147, 150] is summerized in Algorithm (1). Note that Algorithm (1) is not unconditionally stable, since in order for the proximal

Algorithm 1 Primal-Dual Algorithm

1: Initialize $u^0 \in \text{dom } G$ and $p^0 \in \text{dom } F^*$:
2: **while** not converged **do**
3: $u^{k+1} = (I + T^{-1}\partial G)^{-1}\left(u^k - T^{-1}K^T p^k\right)$
4: $p^{k+1} = (I + \Sigma^{-1}\partial F^*)^{-1}\left(p^n + \Sigma^{-1}K(2u^{k+1} - u^k)\right)$
5: $k \leftarrow k + 1$
6: **end while**

algorithm to converge, matrix $\mathcal{M}$ needs to be a positive definite symmetric matrix [150]. This is ensured by the following condition on Σ and T:

$$||\Sigma^{-\frac{1}{2}}KT^{-\frac{1}{2}}|| < 1, \tag{4.41}$$

for a given K. Pock et al. proposed in [150] a family of preconditioners which on one hand allows for large step sizes and on the other still ensures the convergence of the algorithm by fulfilling condition (4.41). This is achieved by setting $T = \text{diag}(\tau)$ and $\Sigma = \text{diag}(\sigma)$ with

$$\tau_j = \frac{1}{\sum_{i=1}^m |K_{i,j}|^{2-\alpha}}, \quad \sigma_i = \frac{1}{\sum_{j=1}^m |K_{i,j}|^{\alpha}}. \tag{4.42}$$

The chosen matrices Σ and T act as a row and column scaling of K. A more sophisticated way of choosing the diagonal scaling factors is devised in [179] where the authors make use of the well-known matrix equilibration algorithm in order to obtain an better diagonal scaling of a matrix. In the remainder of this work we will choose $\alpha = 1$ and use the diagonal preconditioning described in (4.42).

Chapter 5

Discretization

Variational methods are initially defined on a continuous domain. Continuous methods are advantageous because they are more likely to describe real world phenomena and their solutions. The implementation of these methods is flexible in the sense that discretizing continuous methods can be done using different strategies. A link between the discretized problem and the continuous analogue can be established by taking the number of pixels to infinity. This is usually done by analyzing the so called Γ-convergence [28] to the continuous functional. A major challenge in discretizing variational formulations is to ensure that differential operators satisfy the same relations as in the continuous case. A good discretization should also barely depend on the underlying grid. One of the most important differential operators we consider in our applications are the gradient ∇ and its adjoint operator $\nabla^* = -\operatorname{div}$. Their relationship reads as:

$$\int_\Omega \langle \nabla u, p \rangle \, \mathrm{dx} = -\int_\Omega u \operatorname{div} p \, \mathrm{dx} \Longleftrightarrow \nabla^* = -\mathrm{div}, \tag{5.1}$$

which stems from applying integration by part on the left hand side. In order to avoid boundary terms, our discretization needs to fulfill Dirichlet or Neumann boundary conditions. In the following we focus on a simple discretization of the gradient and its matrix-vector representation. This way, we can cast variational methods into the canonical representation (4.13). For the sake of simplicity, we consider the discrete version Ω_h of domain Ω to be the unit square $[0,1] \times [0,1]$ and sample it in a $H \times H$ uniform grid. One possibility to discretize the vector valued gradient operator is the following difference scheme defined for every pixel $(i,j) \in H \times H$ in a forward difference fashion

$$(\nabla u)^1_{i,j} = \begin{cases} u_{i+1,j} - u_{i,j} & \text{if } i < H, \\ 0 & i = H. \end{cases} \tag{5.2}$$

$$(\nabla u)^2_{i,j} = \begin{cases} u_{i,j+1} - u_{i,j} & \text{if } j < H, \\ 0 & j = H. \end{cases} \tag{5.3}$$

Note that the above discretization fulfills the Neumann boundary conditions which are crucial for equation (5.1) to hold. In what follows we devise a gradient matrix operator K

based on the discretization scheme (5.2). We first define a difference matrix ∇_{1D} for the one-dimensional case as follows:

$$K_{1D} = \begin{bmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \\ & & & & 0 \end{bmatrix}. \tag{5.4}$$

It can be easily verified that operator (5.4) calculates forward differences and additionally fulfills the Neumann boundary conditions for a 1D signal. In the following we consider a vectorized version of the 2D image. To this end, we stack the image row-wise and obtain a vector $u \in X$ where $dim\, X = n = H^2$. This allows us to write ∇u as a matrix-vector product. Using the 1D difference matrix K_{1D}, formulating the 2D gradient operator amounts to

$$K = \begin{bmatrix} I_H \otimes \nabla_1 \\ \nabla_1 \otimes I_H \end{bmatrix} \in \mathbb{R}^{2n \times n}, \tag{5.5}$$

where $\otimes$ stands for the Kronecker product and matrix I_H for the H-dimensional identity. Multiplying operator K with a vectorized image results in the following gradient vector:

$$\begin{bmatrix} I_H \otimes \nabla_1 \\ \nabla_1 \otimes I_H \end{bmatrix} u = \begin{pmatrix} K_1 u \\ K_2 u \end{pmatrix}.$$

where K_1 and K_2 stand for the differences in the first and second dimension respectively. Knowing that $\nabla^* = -\text{div}$, calculating the divergence operator is rendered as easy as transposing K since in finite dimensional real vector spaces the adjoint of a linear operator is simply the transpose. We obtain for the divergence operator $-K^T$ the following block matrix

$$\text{div} = -K^T = -\begin{bmatrix} I_H \otimes \nabla_1 \\ \nabla_1 \otimes I_H \end{bmatrix}^T = -\begin{bmatrix} (I_H \otimes \nabla_1)^T & (\nabla_1 \otimes I_H)^T \end{bmatrix} \in \mathbb{R}^{n \times 2n}. \tag{5.6}$$

Chapter 6

Unifying Framework for Semantic Multi-Labeling

6.1 Introduction

The goal of this chapter is to introduce an optimization framework necessary for implementing our works on convex semantic priors [8, 10] included in Chapters **??** and 9. Although different in philosophy, these priors exhibit very similar attributes in terms of realization and optimization. Semantic priors are tailored to favor certain label configurations with disregard to the location, perimeter and neighborhood of the objects. We seek a family of global priors of order $\mathcal{O}(\Omega)$, i.e. every image point is aware of the labels of all other points. In the context of conditional random fields, these priors can be considered as potentials on a fully connected graph. In this chapter, we aim to develop a general method which encodes low-level cues such as color and texture, mid-level cues such as spatial consistency as well as arbitrary semantic priors which can be e.g. drawn from label occurrence statistics. Such an energy can be encoded in model (1.17). Consequently, we can generalize the algorithms in [8, 10, 11] in a prototypical optimization framework. To this end, it is necessary to keep track of the existence of certain label configurations. Specifically one needs to track the existence of individual labels in the segmentation result. For this, we introduce in the next section auxiliary variables l_i which indicate whether a label i is set, i.e. $\exists x \text{ s.t. } u_i(x) = 1$. This yields in a problem formulation which depends on the unknowns u and l. After casting this problem in a saddle point formulation, we devise an generic primal-dual optimization framework for solving general multi-labeling problems with semantic priors. Finally, we apply our framework on a modified version of the classical MDL prior which imposes an upper bound on the label count.

6.2 Label Existence Indicator

In order to track certain label co-occurrences, it is crucial to detect whether individual labels are active in the segmentation result. Building on the formulation in (1.17) subject to (1.19), we introduce an additional variable $l : \mathcal{S} \rightarrow \{0, 1\}^n$ where $\mathcal{S}$ is the convex set introduced in (1.19). In order for l_i to track the existence of a label i, we make the following identification:

$$l_i = \begin{cases} 1, & \text{if } \exists x \in \Omega : \ u_i(x) = 1, \\ 0, & \text{otherwise,} \end{cases} \tag{6.1}$$

Next, we introduce the following constraint:

$$\max_{x \in \Omega} u_i(x) = l_i \qquad \forall i \in \mathcal{L}, \tag{6.2}$$

in order to fulfill the relation in (6.1) which couples u and l. The above constraint will play a major role in devising the semantic priors discussed in this thesis. The global coupling of l is realized by constraint (6.2) which is a non-convex constraint. In order to convexify (6.2), we relax l to the unit interval hence $l : \mathcal{S} \rightarrow [0, 1]^n$ and introduce the following constraint instead:

$$l_i \geq u_i(x) \quad \forall x \in \Omega \quad \forall i \in \mathcal{L}. \tag{6.3}$$

This allows us to write the semantic term E_S in (1.17) with respect to variable l hence $E_S(l)$. Although (6.3) is convex, we need to impose additional restrictions on $E_S(l)$ in order to recover the original coupling (6.2). By showing $E_S(l)$ exhibits the following properties:

1. $E_S(l)$ is convex.
2. $E_S(l)$ is monotonically increasing w.r.t l, i.e.

$$l \preceq l(\bar{u}) \Rightarrow E_S(l) \leq E_S(\tilde{l}), \tag{6.4}$$

we can show that (6.2) is recovered. The proof can be found in a theorem presented in our research paper included in Chapter **??**.

6.3 Saddle Point Formulation

The overall multi-label energy can be summerized as follows:

$$E(u, l) = E_D(u) + E_R(u) + E_S(l) \quad \text{s. t.} \quad u \in \mathcal{S}, \ \ l_i \geq u_i(x), \ \ \forall x \in \Omega, \forall i \in \mathcal{L}. \tag{6.5}$$

We begin by writing out the dataterm E_D and the smoothness term E_S explicitly to obtain the following optimization problem

$$\begin{aligned}\min_{u,l}\max_{p}\sum_{i=1}^{n}\int_{\Omega}u_i(x)\varrho_i(x)\,dx \;+\; \sum_{i=1}^{n}\int_{\Omega}u_i(x)\;\mathrm{div}\,p_i(x)\,dx \\ + E_S(l) + \delta(u\in\mathcal{S}) + \delta(p\in\mathcal{K})\end{aligned}$$

$$\text{s.t. } l_i \geq u_i(x) \quad \forall x\in\Omega, \forall i\in\mathcal{L}, \tag{6.6}$$

where $u \in \mathcal{S}$ denotes the simplex constraint defined in (1.19) and $p \in \mathcal{K}$ the constraint on the dual variable. Depending on the type of div operator and constraint $\mathcal{K}$ we achieve different types of regularizations. These regularizations can be of local [37, 206] or non-local [199] nature. In contrast to the constraints $p \in \mathcal{K}, u \in \mathcal{S}$ which can be handled using a closed form resolvent, the inequality $l_i \geq u_i(x)$ does not exhibit an orthogonal projection scheme. Hence, we reformulate our problem by introducing Lagrangian multiplier θ which yields the following modified problem:

$$\begin{aligned}\min_{u,l}\max_{p,\theta}\sum_{i=1}^{n}\int_{\Omega}u_i(x)\varrho_i(x)\,dx \;+\; \sum_{i=1}^{n}\int_{\Omega}u_i(x)\;\mathrm{div}\,p_i(x)\,dx + \sum_{i=1}^{n}\int_{\Omega}\theta(x)(u_i(x)-l_i) \\ + E_S(l) + \delta(u\in\mathcal{S}) + \delta(p\in\mathcal{K}) + \delta(\theta\preceq 0).\end{aligned} \tag{6.7}$$

This min-max formulation is amenable to the primal-dual solver introduced in Algorithm 1, provided $E_S(l)$ is a "simple" function in the sense that its resolvent (4.40) can be computed in closed-form.

6.4 Primal-Dual Algorithm for Multi-Label Segmentation with Semantic Priors

Following the setting in Chapter 4, we discretize the problem by setting $u, \varrho, \theta \in X^n$ and $p \in Y^n$ after vectorizing the image domain. From now on, the divergence operator div is considered to be a matrix. We identify the stacked variables $(u, l)^T$ as the primal unknown. The dual variable is composed of the stacked variables $(p, \theta)^T$. The most important step to solve an optimization problem using Algorithm 1 is recognizing the linear operator K and the functions G and F^* in the saddle point formulation (6.7). This allows us to map

formulation (6.7) to the primal-dual formulation (4.19) as follows:

$$\left\langle K \begin{pmatrix} p \\ \theta \end{pmatrix}, \begin{pmatrix} u \\ l \end{pmatrix} \right\rangle = \sum_{i=1}^{n} \langle u_i, \operatorname{div} p_i \rangle + \sum_{i=1}^{n} \langle \theta_i, (u_i - l_i) \rangle \tag{6.8}$$

$$G \begin{pmatrix} u \\ l \end{pmatrix} = \langle u, \varrho \rangle + \delta(u \in \mathcal{S}) + E_S(l) \tag{6.9}$$

$$F^* \begin{pmatrix} p \\ \theta \end{pmatrix} = \delta(p \in \mathcal{K}) + \delta(\theta \succeq 0). \tag{6.10}$$

By applying Algorithm 1, we obtain the update steps listed in Algorithm 2.

Algorithm 2 Primal-Dual Algorithm for Multi-Label Segmentation with Semantic Priors

1: **while** not converged **do**
2: $u_i^{k+\frac{1}{2}} = u_i^k - \tau_1 \left(\operatorname{div} p_i + \theta_i \right)$
3: $l_i^{k+\frac{1}{2}} = l_i^k - \tau_2 \sum_x \theta_i(x)$
4: $\begin{pmatrix} u^{k+1} \\ l^{k+1} \end{pmatrix} = (I + T^{-1} \partial G)^{-1} \begin{pmatrix} u^{k+\frac{1}{2}} \\ l^{k+\frac{1}{2}} \end{pmatrix}$
5: $p_i^{k+\frac{1}{2}} = p_i^k - \sigma_1 \nabla \left(2u_i^{k+1} - u_i^k \right)$
6: $\theta_i^{k+\frac{1}{2}} = \theta_i^k + \sigma_2 \left(\left(2u_i^{k+1} - u_i^k \right) - \left(2l_i^{k+1} - l_i^k \right) \right)$
7: $\begin{pmatrix} p^{k+1} \\ \theta^{k+1} \end{pmatrix} = (I + \Sigma^{-1} \partial F^*)^{-1} \begin{pmatrix} p^{k+\frac{1}{2}} \\ \theta^{k+\frac{1}{2}} \end{pmatrix}$
8: $k \leftarrow k + 1$
9: **end while**

$\Sigma^{-1} = diag((\sigma_1, \sigma_2)^T)$ and $T^{-1} = diag((\tau_1, \tau_2)^T)$ are diagonal preconditioning matrices which are set according to the strategy described in (4.42). The resolvent operators are applied component-wise and consist of

$$(I + T^{-1} \partial G)^{-1} \begin{pmatrix} u \\ l \end{pmatrix} = \begin{pmatrix} \Pi_{\mathcal{S}}(u - \tau_1 \varrho) \\ (I + \tau_2 \partial E_S)^{-1}(l) \end{pmatrix} \tag{6.11}$$

and

$$(I + \Sigma^{-1} \partial F^*)^{-1} \begin{pmatrix} p \\ \theta \end{pmatrix} = \begin{pmatrix} \Pi_{\mathcal{K}}(p) \\ \Pi_{\mathbb{R}_+}(\theta) \end{pmatrix}. \tag{6.12}$$

The orthogonal projection operator $\Pi_{\mathcal{S}}(u)$ can be calculated using the simplex projection algorithm described in [126] which uses at most n steps in order to converge. The calculation of the projector $\Pi_{\mathbb{R}_+}(\theta)$ can be realized using a simple point-wise clipping into the positive numbers, i.e.

$$\Pi_{\mathbb{R}_+}(\theta) = \max(0, \theta), \tag{6.13}$$

whereas the projection $\Pi_{\mathcal{K}}(u)$ hinges on the type of relaxation of the Potts model and vary from simple point-wise projections as in the case of the Zach relaxation [206] to a

more complicated projection technique as in the CCP method [37] which has a quadratic complexity.

If E_s is simple, multi-labeling algorithms with semantic priors only differ in the resolvent operator

$$(I + \tau_2 \partial E_S)^{-1}. \tag{6.14}$$

In case (6.14) is not trivial to calculate as in [8, 10], an introduction of auxiliary variables and the modification of the bilinear term (6.8) are sufficient to render all resolvents simple.

6.5 Implementation on Parallel Architectures

First order optimization methods like Algorithm 1 are ideal to be implemented on parallel architectures. This is due to the simple update steps which consist of applying a matrix-vector product and simple pointwise operations. Typically matrix K is sparse and is of known structure. This makes it possible to calculate the matrix-vector product $\langle K, \cdot \rangle$ point-wise making a parallel implementation straightfoward. However, in the case of the semantic primal dual Algorithm 2, matrix K contains dense rows corresponding to the l update. Consequently, this coupling requires performing a summation over θ. To this end, we exploit parallel architecture by applying a reduction scheme described in [82].

6.6 Generalized MDL Prior

In this section, we explore an example of a semantic prior on which Algorithm 2 is applicable. One of the simplest strategies of penelizing label configurations is imposing an MDL prior as described in Section 1.4.2. This means that we encourage a parsimonious description using fewer labels. In [11] we presented a family of MDL priors which generalize the usual linear penalty of the label count.The linear MDL penalty can be written as:

$$E_S(l) = \sum_i^n l_i C_i, \tag{6.15}$$

where C_i represents a predefined cost for each label occurring in the image. This linear prior has been previously considered in [51, 204]. We extend (6.15) to

$$E_S(l) = f\left(\sum_{i=1}^n l_i\right), \tag{6.16}$$

with $f : \mathbb{R} \to \mathbb{R}$ being an arbitrary convex and monotonously increasing function. Specific choices are $f(s) = s$ as in [51, 204], $f(s) = s^2$, or $f(s) = \delta(s \leq s_0)$ for some $s_0 \geq 1$, i.e

$$f(s) = \begin{cases} 0, & s \leq s_0, \\ \infty, & \text{else}, \end{cases} \tag{6.17}$$

The latter function imposes a hard constraint on the total number of active labels. We introduced this prior in a technical report [11]. The constraint in (6.17) allows us to optimize over regions when knowing the maximal number of labels beforehand. A straightforward application of this constraint is devising a convex relaxation formulation of the piece-wise constant Mumford-Shah model [131] as we will demonstrate in the next section.

Convex Relaxation of the Piece-Wise Constant Mumford-Shah Model

In the previous section, we reviewed the linear MDL prior and generalized it to a composition with convex, monotonously increasing functions f. In this section we choose $f(s) = \delta_{s \leq s_0}$ and obtain

$$E_S(l) = \delta\left(\left(\sum_i^n l_{i=1}\right) \leq s_0\right), \tag{6.18}$$

which limits the label count to at most s_0. For the dataterm we choose

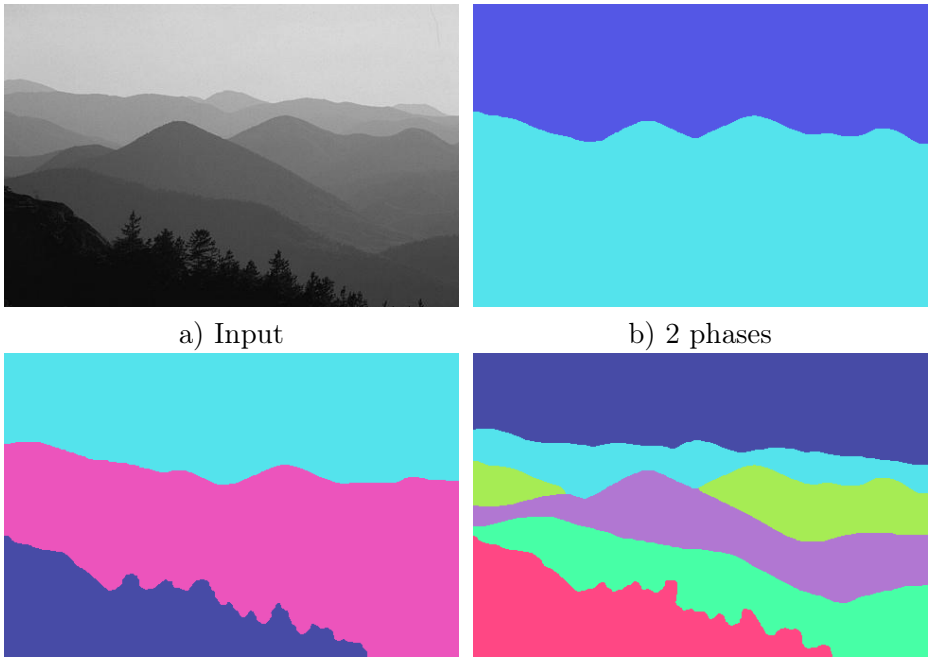
$$\varrho_i(x) = (I(x) - c_i)^2. \tag{6.19}$$

By setting the number of candidate labels to the number of gray values in a 8 bit image $n = 256$ and $c_i = i$, $0 \leq i \leq 255$, we obtain a convex relaxation of the piece-wise constant Mumford-Shah functional [131]. This allows for the simultaneous optimization for the regions Ω_i and implicitly for the mean gray values c_i. Solving for the means and the clusters for a fixed number of labels reminds us of k-means clustering. In contrast to k-means, our algorithm allows for controlling the smoothness of the regions which allows for better handling of noisy input data. Additionally, our the TV smoothness can be made edge [29] aware or even non-local [199]. Finally, our method relies on a globally optimal optimization scheme and does not resort to alternating minimizing. However, global optimality comes with the price of having to solve a huge optimization problem since we discretize the range of the input image.

6.7 Implementation

Since E_S in (6.18) is simple, we can apply Algorithm 2 to solve problem (6.7). To implement the algorithm, we need to account for performing a resolvent step on the term $E_S(l)$ in (6.18) in each iteration. The simplicity of (6.18) allows calculating its resolvent in a closed which reads as:

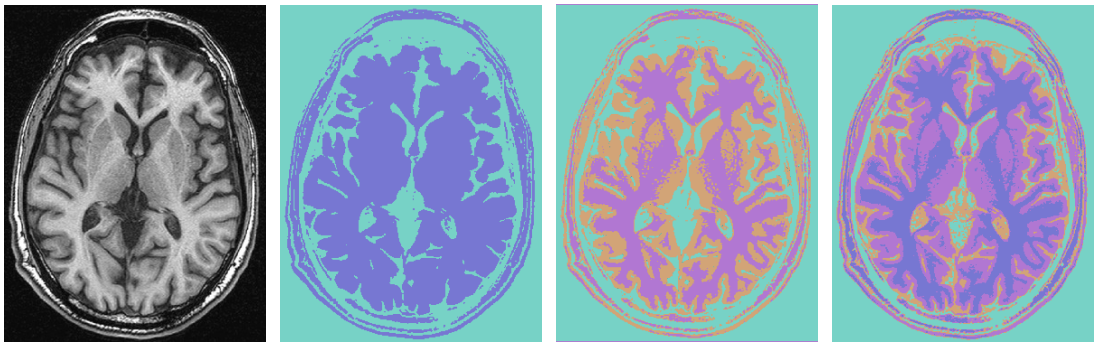
$$l = (I + \tau_2 \partial E_S)^{-1}(\tilde{l}) \leftrightarrow \tilde{l}_i - \frac{1}{n} \max\left(s_0 - \sum_{i=1}^n \tilde{l}_i, 0\right). \tag{6.20}$$



a) Input b) 2 phases

c) 3 phases d) 6 phases

Figure 6.1: We apply our label count constraint on a mountain scene a) and obtain different segmentation results and mean values c_i depending on the selected value for s_i: b) $s_i = 2$ c) $s_i = 3$ d) $s_i = 6$.



a) Input b) 2 phases c) 3 phases d) 4 phases

Figure 6.2: The generalized MDL prior allows imposing an upper bound on the number of labels. This example shows an unsupervised segmentation of MRI image (a) into 2 regions ($s_0 = 2$) (b) 3 regions ($s_0 = 3$) (c) and 4 regions ($s_0 = 4$) (d).

6.8 Experiments

The piece-wise constant Mumford-Shah model can be applied in various fields like medical imaging or approximating an image with a piece-wise constant function in order to obtain a reasonable simplification. The image formation process in medical imaging is prone to severe noise and wrong readings. This makes interpreting data originating from medical scanners a tedious process and a post processing step becomes crucial in order for the doctor to properly interpret scans and conclude a diagnosis. Detecting matters (white, gray) in magnetic resonance imaging (MRI) is a very important task. The piece-wise constant Mumford-Shah model is an excellent approximating model since each region typically exhibits a common mean gray value. Our algorithm is able to jointly dissect different regions Ω_i in MRI and implicitly estimate their associated mean values c_i. In Figure 6.2 we apply our method on an MRI brain scan for different bounds s_0 on the label count. In Figure 6.1 we apply our label count constraint on a mountain scene and obtain different segmentation results depending on the selected value for s_0.

6.9 Conclusion

In this chapter, we proposed a unifying framework for variational multi-label segmentation with semantic priors. To this end, we devised a generic primal-dual algorithm which serves as a guide to solving multi-label problems with arbitrary convex semantic priors. We applied our framework on a modified version of the MDL prior in which we impose an upper bound on the label count. This allows us to propose a convex relaxation of the piece-wise Mumford-Shah model. This chapter serves as a generalization of our published works on semantic priors which are included in Chapters **??** and 9

Chapter 7

Entropy Minimization for Mixed-Integer Programs

7.1 Introduction

In recent years, label assignment problems became a major research field in computer vision and machine learning. These problems can vary depending on the application, however conceptionally they are similar. Assigning a label can be described by an optimization problem. Since one typically deals with a discrete label space $\mathcal{L} \subset \mathbb{Z}$, we refer to these optimization problems as integer programs (IP). In this section, we deal with a broader class of problems namely mixed-integer programs (MIPs). The unknowns in MIPs are partly real valued and partly constrained in $\mathcal{L}$. The resulting generic optimization problem is given by:

$$\begin{aligned} \min \;& F(\varphi) \\ \text{s.t.} \;& \varphi \in \mathcal{C} \\ & \varphi(x_{\mathcal{Z}}) \in \mathcal{L}, \; \forall \; x_z \in \mathcal{I}, \end{aligned} \tag{7.1}$$

where $\mathcal{I} \subseteq \Omega$ denotes a subset of domain $\Omega_0 \subset \mathbb{R}^N$ in which φ takes on values in $\mathcal{L}$. The constraint $\mathcal{C} \subset \Omega$ and the objective $F : \Omega_0 \to \mathbb{R}$ are convex. Overall, formulation (7.1) is not convex due to the integer constraints $\varphi(x_z) \in \mathcal{L}$. We denote the problem as an integer programming problem if $\mathcal{I} = \Omega_0$. Most MIPs can be be simplified in terms of search space, meaning that we can find an equivalent formulation which is $\{0, 1\}$ valued [195]. The resulting mixed-integer programming problem looks as follows:

$$\begin{aligned} \min \;& E(\xi) \\ \text{s.t.} \;& \xi \in C \\ & \xi(x_b) \in \{0, 1\}, \; \forall \; x_b \in \mathcal{B}, \end{aligned} \tag{7.2}$$

where function $E : \Omega \to \mathbb{R}$ and constraint $C \subset \Omega$ are convex as in the original MIP (7.1) and $\mathcal{B} \subseteq \Omega$ denotes the index set where ξ takes on binary values. The domain Ω is typically a

lifted version of Ω_0. For example, this can be realized by setting $\Omega = \Omega_0 \times \mathcal{L}$, e.g. in [38, 147, 149]. Hence the simpler binary constraint comes at the cost of an optmization problem with a higher dimension. Problem (7.2) is hard to solve since its domain remains non-convex. Hence classical optimization methods are unfeasible. As a remedy, an optimization over the interval $[0, 1]$ is usually performed and the resulting convex problem reads:

$$\begin{aligned} \min \;& E(\xi) \\ \text{s.t.} \;& \xi \in C, \\ & \xi(x_b) \in [0, 1], \; \forall \; x_b \in \mathcal{B}. \end{aligned} \tag{7.3}$$

Note that in case $E(\xi)$ is linear and C is polyhedral, the problem in (7.3) is referred to as the LP-relaxation of (7.2). Although we can obtain an optimal solution to (7.3), this does not guarantee to solve the original binary problem (7.2). A remarkable exception is a special LP case where the underlying constraint matrix is totally uni-modular [85, 164] for which it was proven that the results of the relaxed formulation solve the original binary problem. A special case of LPs with totally uni-modular matrices is the class of sub-modular functions which can be minimized in polynomial time using graph-cut algorithms [25]. This fact is also reflected in the LP relaxation of sub-modular functions which is proven to be tight [200]. In the general LP case, where the constraint matrix is not totally uni-modular, the solutions are usually not integral and a posteriori rounding schemes such as branch-and-bound strategies or the cutting-plane method can be used to recover an integer solution. However, although very effective in practice, these methods are not applicable for general MIPs. For a detailed study we encourage the reader to look into [133]. In case $E(\xi)$ or C are non-linear, it is not obvious when the relaxed problem recovers the original problem. A notable exception is an algorithm tackling the two-phase segmentation task [138] where one can prove that every threshold of the relaxed solution is a minimizer of the binary problem. However, this requires 1-homogeneity of the objective function.

7.2 Shannon's Entropy as a Measure of Tightness

In computer vision, minimizing Shannon's entropy has been already applied to shadow removal [63] and unsupervised image segmentation [180]. In Chapter, **??** we included our research paper [9] in which we are the first to propose incorporating the entropy of the objective variable as a measure of its integrality. We are able to show in theory as well as experimentally that by extending the relaxed multi-label problem by an entropy term, we can improve on the integrality of the solution. Consequently, we obtain tighter solutions compared to the relaxed convex problem (7.3) with respect to the original binary problem. The entropy function for a $[0, 1]$ distribution is illustrated in Figure 7.1 where it becomes immediately obvious why the entropy encourages more binary solutions. Shannon's entropy is applied point-wise in order to penalize deviations from the binary constraint $\xi(x_b) \in \{0, 1\}$ in (7.2). The total amount of entropy can be formulated as follows:

$$H(\xi) = -\int_{\mathcal{B}} (\xi(x) \log \xi(x)) + ((1 - \xi(x)) \log(1 - \xi(x))) \; dx. \tag{7.4}$$

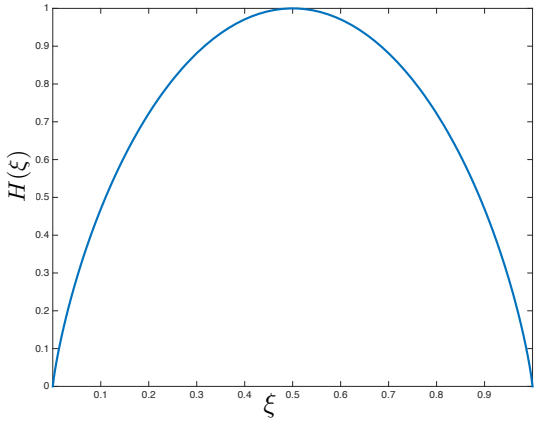


Figure 7.1: The entropy of ξ peaks ($H(\xi) = 1$) at maximal uncertainty $\xi = 0.5$ and is minimal for $\xi = 0$ and $\xi = 1$.

Augmenting problem (7.3) with Shannon's entropy (7.4) results in the following optimization problem:

$$\begin{aligned} &\min \;\; E(\xi) + \theta H(\xi) \\ &\text{s.t.} \;\; \xi(x_b) \in [0,1], \;\; \forall \; x_b \in \mathcal{B}, \end{aligned} \tag{7.5}$$

where parameter $\theta \geq 0$ is a scalar allowing to fine tune the soft constraint $H(\xi)$. Note that by setting $\theta = \infty$ we recover the orginal binary problem (7.2). Hence, θ relates both the binary MIP and the relaxed MIP. Formulation (7.5) can be considered a generalization of our method in [9] since in this chapter we handle the case of the broader class of MIPs. Unfortunately, the concavity of the entropy function renders (7.5) non-convex. Luckily, the problem formulation has a special, well-studied structure: it can be written as a difference of convex functions (DC). For this class of problems researchers devised a family of algorithms commonly known as DC-Programming. We will elaborate on DC-Programming in the next section.

7.3 DC-Programming

In [182] Tao et al. devised an algorithm specialized in solving a family of non-convex problems of the following form:

$$\alpha = \inf_{\xi} \{g(\xi) - h(\xi)\}, \tag{7.6}$$

where g and h are proper lower semi-continuous convex functions. One of the most important insights for minimizing problem (7.6) is the fact that its dual problem

$$\beta = \inf_{\xi} \{h^*(\xi) - g^*(\xi)\}, \tag{7.7}$$

is symmetric [184] with respect to Fenchel conjugacy and sign. Furthermore, it is important to mention that $\alpha = \beta$ holds. DC-Programming is closely related to majorization-minimization [105] and has been applied in vision to quadratic program relaxations of MAP inference [91]. A similar well-known approach, allthough limited to differentiable functions, is referred to as the convex-concave procedure (CCCP) [205]. In [182], the authors are able to derive a simple iterative procedure for solving (7.6) based on the dual (7.7) and the KKT systems of (7.6) and (7.7). The overall abstract DC algorithm performs the iterative steps illustrated in Algorithm 3 and is guaranteed to converge to a critical point.

Algorithm 3 The DC algorithm

Initialize $\xi^0 \in \operatorname{dom} g, \varphi^0 \in \operatorname{dom} h$.
while not converged **do**
  Choose $\varphi^k \in \partial h(\xi^k)$
  Choose $\xi^{k+1} \in \partial g^*(\varphi^k)$
  $k \leftarrow k + 1$
end while

The second step in Algorithm 3 can be simplified by using Fenchel duality

$$\xi^{k+1} \in \partial g^*(\varphi^k) \Leftrightarrow \varphi^k \in \partial g(\xi^{k+1}) \Leftrightarrow 0 \in \partial g(\xi^{k+1}) - \varphi^k. \tag{7.8}$$

Since $0 \in \partial g(\xi^{k+1}) - \varphi^k$ is the optimality condition of the following convex function:

$$\xi^{k+1} = \arg\min_{\xi} g(\xi) - \langle \varphi^k, \xi \rangle, \tag{7.9}$$

we end up solving the monotone inclusion $\xi^{k+1} \in \partial g^*(\varphi^k)$ in each iteration by solving (7.9) using a solver of our choice. Intuitively Algorithm 3 performs a linearization of $h(\xi)$ function in each step. Thus, it can be considered a majorization-minimization technique [105] using a linear majorization of the concave part in (7.6).

7.4 Minimizing the L_0 Norm for Image Cartooning

Our algorithmic framework (7.5) can be utilized for any instance of a mixed-integer program such as production planning, scheduling, optimization of telecommunication networks and some instances of imaging problems. In this thesis, we consider an example from computer vision, namely image cartooning using L_0 smoothing. For this, it is important to name a few related works. The first known approach to directly minimizing the L_0 norm of the gradient of the objective variable is the work by Xu et al. [201]. The authors use a quadratic decoupling minimization in order to tackle the non-convex problem and increase the coupling parameter in a continuation fashion as has been done in the context of optical flow optimization [169]. However, it is not known wether the minimization procedure solves the original problem. Another approach for computing image cartoons is the real-time piece-wise smooth method by Strekalovskiy et at. [173] where the authors devise a non-convex variant of the primal-dual Algorithm 1 by modifying the resolvent operator via

Moreau's identity [157]. However, these methods are specialized algorithms and are not trivially applicable to other MIPs. In this section, we make use of the entropy in order to solve a mixed-integer formulation of the L_0 penalty in the context of image cartooning. By applying this formulation on the gradient of natural images, it is possible to impose a piece-wise constant prior on the resulting image. Along with a data fidelity term, we obtain the following cost function:

$$E(u) = \frac{\lambda}{2} \|u - f\|_2^2 + \|\nabla u\|_0 \,, \tag{7.10}$$

where the unknown is $u : \Omega \to \mathbb{R}^c$ and the input image is $f : \Omega \to \mathbb{R}^c$ with c being the number of channels. The term $\|\nabla u(x)\|_0$ is a sparsity measure of the gradient, i.e.

$$\|\nabla u\|_0 = \#\{x : \|\nabla u(x)\|_F \neq 0\}. \tag{7.11}$$

The term $\|\nabla u(x)\|_F$ denotes the Frobenius norm of the gradient matrix and covers gray-valued as well as multi-channel images. It is known in the compressed sensing community that problem (7.10) is NP-hard [35]. The most popular remedy is to replace the $\|\nabla u\|_0$ norm by a L_1 regularization of the gradient, leading to the following convex problem:

$$E(u) = \frac{\lambda}{2} \|u - f\|_2^2 + \|\nabla u\|_1 \,, \tag{7.12}$$

where

$$\|\nabla u\|_1 = \int_\Omega \|\nabla u(x)\|_F \,\mathrm{dx} \,. \tag{7.13}$$

It turns out that model (7.10) is exactly the Rudin–Osher–Fatemi (ROF) model introduced in [161] which has been extended to the vectorial case in [53, 72]. As a consequence of the co-area formula (1.9), the total variation effectively measures the perimeter of all level sets [41]. This effect invokes a loss in contrast which can be observed in the reconstructions. Next, we rewrite problem (7.10) by replacing the L_0 norm by a mixed-integer formulation introduced in [66] which gives

$$\begin{aligned} \min_{u,\alpha} \quad & \frac{\lambda}{2} \|u - f\|_2^2 + \|\alpha\|_1 \\ \text{s. t.} \quad & \alpha(x) M \geq \|\nabla u(x)\|_F \,, \\ & \alpha(x) \in \{0, 1\}, \\ & \forall x \in \Omega, \end{aligned} \tag{7.14}$$

with $\alpha : \Omega \to \{0, 1\}$ and $M \in \mathbb{R}_+$ being a scalar which needs to be provided beforehand, indicating a maximal jump in image values. Instead of a problem with a non-convex objective function as in (7.10) we end up with a convex problem with the non-convex integer constraint $\alpha(x) \in \{0, 1\}$. Similar to (7.5), we relax this constraint to $\alpha(x) \in [0, 1]$ and introduce an entropy term in order to encourage the integrality of α. By doing so, we

obtain

$$\begin{aligned} \min_{u,\alpha} \quad & \frac{\lambda}{2} \|u - f\|_2^2 + \|\alpha\|_1 + \theta H(\alpha) \\ \text{s.t.} \quad & \alpha(x) M \geq \|\nabla u(x)\|_F , \\ & \alpha(x) \in [0,1], \\ & \forall x \in \Omega. \end{aligned} \tag{7.15}$$

Although problem (7.15) can be cast in a canonical DC-Programming form as in (7.6), it is not obvious how to solve the subproblems (7.9). This is due to the non-linear constraint $\alpha(x)M \geq \|\nabla u(x)\|_F$. By introducing an auxiliary variable p we are able to transform this constraint into an second order cone (SOC) and we find

$$\begin{aligned} \min_{u,\alpha,p} \quad & \frac{\lambda}{2} \|u - f\|_2^2 + \|\alpha\|_1 + \theta H(\alpha) \\ \text{s.t.} \quad & \alpha(x) M \geq \|p\|_F , \\ & p = \nabla u, \\ & \alpha(x) \in [0,1] \\ & \forall x \in \Omega. \end{aligned} \tag{7.16}$$

Calculating the inclusion (7.9) amounts to solving the following problem:

$$\begin{aligned} \arg\min_{u,\alpha,p,l} \max_{\beta,\zeta} \quad & \frac{\lambda}{2} \|u - f\|_2^2 + \|\alpha\|_1 + \langle \varphi^k, \alpha \rangle \\ & + \langle \beta, p - \nabla u \rangle + \langle \zeta, l - M\alpha \rangle, \\ \text{s.t.} \quad & l(x) \geq \|p\|_F , \\ & \alpha(x) \in [0,1], \\ & \forall x \in \Omega, \end{aligned} \tag{7.17}$$

with $\varphi^k \in \partial(-\theta H(\alpha^k))$. Note that we introduced an auxiliary variable l in order to decouple the the $[0,1]$ constraint from the SOC constraint and to avoid applying alternating projections as in [27]. Additionally, Lagrange multipliers β and ζ are introduced in order to handle the constraints $p = \nabla u$ and $l = M\alpha$ respectively. Formulation (7.17) is a saddle point problem which we can solve using Algorithm 1. For this, we make the

identifications

$$\left\langle K \begin{pmatrix} u \\ \alpha \\ p \\ l \end{pmatrix}, \begin{pmatrix} \beta \\ \zeta \end{pmatrix} \right\rangle = \langle \beta, p - \nabla u \rangle + \langle \zeta, l - M\alpha \rangle$$

$$G \begin{pmatrix} u \\ \alpha \\ p \\ l \end{pmatrix} = \frac{\lambda}{2} \|u - f\|_2^2 + \|\alpha\|_1 + \langle \varphi^k, \alpha \rangle + \delta \left(l(x) \geq \|p(x)\|_F, \forall x \in \Omega \right)$$

$$+ \delta \left(\alpha(x) \in [0,1]^n, \forall x \in \Omega \right)$$

$$F^* \begin{pmatrix} \beta \\ \zeta \end{pmatrix} = 0. \tag{7.18}$$

For the sake of simplicity, the right hand sides are given, by abuse of notation, in the continuous setting. A discretization can be realized trivially, following Chapter 5. In order to perform the necessary steps in Algorithm 1, we need to compute the resolvent operator of G which is given component-wise in the discrete setting:

$$(I + T^{-1}\partial G)^{-1} \begin{pmatrix} u \\ \alpha \\ p \\ l \end{pmatrix} = \begin{pmatrix} \frac{u + \tau_1 \lambda f}{1 + \tau_1} \\ \Pi_{[0,1]}(\hat{\alpha}) \\ \Pi_{\mathrm{SOC}}(p, l) \end{pmatrix}, \tag{7.19}$$

where $\hat{\alpha} = \alpha + \tau_2(1 + \varphi^k)$. The diagonal preconditioning matrices $T^{-1} = diag((\tau_1, \tau_2, \tau_3, \tau_4)^T)$ and $\Sigma^{-1} = diag((\sigma_1, \sigma_2)^T)$ are set according to (4.42). The projection $\Pi_{[0,1]}(l)$ is a simple clipping to the $[0, 1]$ interval, i.e.

$$\Pi_{[0,1]}(l) = \max(\min(l, 1), 0). \tag{7.20}$$

The point-wise SOC projection $\Pi_{\mathrm{SOC}}(\alpha, p)$ can be computed in closed-form [145] . The overall DC procedure is summarized in Algorithm 4.

Algorithm 4 DC-Cartooning Algorithm

Initialize $u^0, \alpha^0, p^0, l^0, \beta^0, \zeta^0$.
while not converged **do**
  Choose $\varphi^k \in \partial(-\theta H(\alpha^k))$
  Solve (7.17) to obtain $(u^{k+1}, \alpha^{k+1}, p^{k+1}, l^{k+1}, \beta^{k+1}, \zeta^{k+1})$
  $k \leftarrow k + 1$
end while

7.5 Experiments

In the following section, we demonstrate the effectiveness of our MIP approach in its ability to minimize problem (7.16). For this, we use natural images and estimate their

a) Input b) $\lambda = 6$ c) $\lambda = 3$ d) $\lambda = 2$

Figure 7.2: A cartoon approximation of a natural image using different smoothness parameters λ. All results were computed with $\theta = 1$.

cartoon approximations. Additionally, we reconstruct cartoon images contaminated with JPEG compression artifacts. We demonstrate that penalizing the L_0 norm of the image gradient is a suitable prior for such task. We also compare our approach with both the L_0 smoothing method in [201] and the piece-wise constant Mumford-Shah approach in [173]. In Figure 7.2, a cartoon approximation of a natural image is calculated using different smoothness parameters λ. With decreasing λ our method penalizes jumps more strictly.

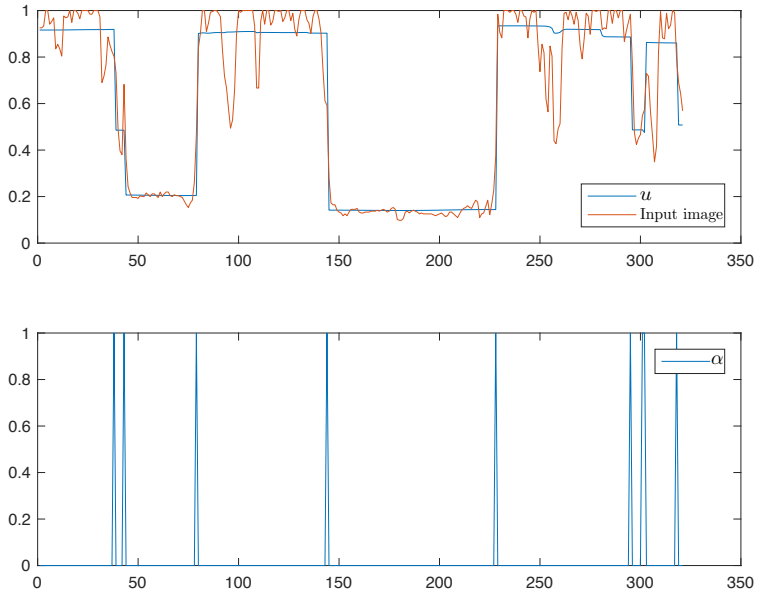


Figure 7.3: A cross section of the L_0 reconstruction (upper plot) of the input image (red channel) in Figure 7.2 along with variable α (lower plot) which tracks the jumpset $\{x : \|\nabla u(x)\|_F \neq 0\}$. Note how the peaks in α exactly coincide with the jumps in u.

a) Input b) Cropped input c) ROF, PSNR=23.69

d) [201], PSNR=24.65 e) [173], PSRN=25.4366 f) Ours, PSRN=25.4395

Figure 7.4: Image deblocking of image in (a) using ROF denoising (b), the method by Xu et al. [201] (d), the real-time Mumford-Shah approach of [173], and our formulation (7.15). Our method is able to achieve the best smoothing result while still recovering fine structures.

Figure 7.3 shows a cross section of an approximation with $\lambda = 3$. In this 1D plot, we observe that our method produces sparse gradients while minimizing the dataterm. In Figure 7.4 we apply our algorithm on a cartoon image corrupted with JPEG compression artifacts. We compare our algorithm to the ROF model ($\theta = 0$) and to the methods of Xu et al. [201] and Strekalovskiy et al. [173]. We tuned all parameters such that we obtain the best reconstructions in terms of PSNR values. In contrast to ROF denoising, our method clearly exhibits superior performance at de-blocking the image and recovering fine structures of the input image. When compared to [201] and [173], we observe that our method achieves visually better reconstruction results than [201] and a comparable result to the method of Strekalovskiy et al. in terms of PSRN values. However, in contrast to [173], our method is guaranteed to converge due to the provably convergent DC-Algorithm 4. Note that in all our experiments we initialize all variables with zero. A dependence of the reconstruction results on the initialization cannot be observed.

7.6 Conclusion

In this chapter, we generalized our work on introducing entropy to continuous multi-label optimization [9] (included in Chapter 10). To this end, we discussed the broader class of MIPs and augmented their convex relaxation with an entropy term. We demonstrate the effectiveness of our approach by tackling the problem of image cartooning which can be formulated as an MIP. Our experimental results show that our approach competes with state of the art specialized cartooning algorithms. Finally, our algorithm is guaranteed to converge because we make use of a provably convergent DC algorithm.

Part II

A Selection of Own Publications

Chapter 8

Paper Summaries

In the following, we provide summaries for the research papers included in this part.

Convex Optimization for Scene Understanding [8]

The most prominent strategies for semantic priors are based on pre-learning co-occurrence statistics and penalize label configurations which are not likely to occur. These priors have been incorporated in a discrete multi-label approach by Ladicky et al. [104] and have been introduced to the spatially continuous setting in our work in [10]. Co-occurrence priors always depend on a training set and are thus prone to overfitting. In order to avoid this, a more principled way of incorporating context has been introduced by Delong et al. in [50], namely hierarchical priors. These priors group labels in a context hierarchy. To this end, labels are grouped as leafs in a tree where parent nodes represent scene labels. This natural approach is able to refine image multi-labeling in a more intuitive and principled way. Additionally, it is less prone to overfit to a database. This comes from the fact that grouping labels under a context is a universal principle and comes close to the human intuition. In this work we propose incorporating hierarchical priors into continuous multi-label optimization. Our algorithm is able to jointly infer the labeling of the scene as well as the scene type. For example, the labels 'car', 'road', 'building' help determining the scene type 'urban' and in contrary knowing the context e.g. urban, nature, outdoor and indoor helps refine the pixel-wise labels. In contrast to [50], we devise an algorithm which does not depend on the underlying grid and thus does not introduce metrication errors [95, 137]. We cast the hierarchical multi-label algorithm into one single optimization problem in contrast to the discrete approach [50] which performs sequential optimization. Our algorithm is parallelizable and trivially extendable to 3D multi-labeling unlike discrete graph-cut based approaches which do not scale well with increasing dimensions [95].

Entropy Minimization for Convex Relaxation Approaches [9]

Formally label assignment problems map from some domain to a discrete label space. They can be formulated as integer optimization problems. These integer problems are typically formulated by means of a $\{0, 1\}$ valued indicator functions. A common strategy to solve these NP-hard problems is to relax their range to the interval $[0, 1]$. The resulting convex relaxation formulations can be solved using state-of-the-art convex optimization solvers. However, the minimizers of these methods are not always binary and guaranties on the quality of the solution with respect to the original binary problem cannot be made a priori. Typically, a binary solution is recovered by a suboptimal a posteriori thresholding technique. In this work, we introduce a binarization energy to the optimization problem and enforce the solution to become less fractional. To this end, we incorporate Shanon's entropy as a measure of tightness. Although the resulting cost function is not convex, it happens to be is a difference of convex (DC) functions which can be minimized using a specialized algorithm. This so called DC algorithm does not recover the global solution, however, its convergences to a critical point is provably guaranteed. In our work we show theoretically and experimentally that incorporating the entropy consistently improves the multi-labeling results. Applications range from multi-label inpainting to spatio-temporal 3D reconstruction. By measuring the a posteriori optimality gap we are able to show that our solutions exhibit tighter energies than classical convex relaxation approaches. In terms of computational time our methods do not change the complexity of the algorithm. Indeed, our method speeds up the convergence behavior of convex relaxation methods by a factor of two.

Smooth Piece-Wise Rigid Scene Flow from RGB-D Images [5]

Estimating motion is one of the most important tasks in human and machine vision. Its application can be as obvious as predicting moving objects in a dynamic scene as well as generating super-resolved frames in time [77] and space [188]. While 2D motion estimation approaches have been around since the 80's [121], in recent years methods for estimating 3D motion are getting more popular. This is due to the availability of high quality RGB-D sensor's and massively parallel architectures such as GPU's which are available on end-user graphics cards. Additionally, there has been a shift in convex optimization algorithms towards first-order methods. These iterative methods can be implemented in parallel and often lead to real time performance. Thus solving large scale variational problems efficiently such as image segmentation, becomes feasible. Another trending topic in computer vision and robotics is visual odometry, where the challenge is to predict camera poses from a sequence of RGB [55] or RGB-D [94] frames. Traditional scene flow techniques are based on linearizing the optical and range flow [197] constraints and solve for the displacement between corresponding points. In this paper, we tackle the problem of jointly segmenting the scene into rigidly moving parts and estimating their motion. To this end, we borough techniques from image multi-labeling and visual odometry and devise an energy function which is dependent on the partitioning of the scene and its underlying motions. Solving these highly correlated problems jointly helps solving ambiguities in motion and in image segments. In addition to rigidly moving parts, our method is able to recover non-rigidly moving points by allowing a smooth transition between labels. This is realized by allowing the labeling to vary between 0 and 1 as well as choosing a simple regularization which promotes a smooth transition between labels when necessary. Detecting non-rigidity in the scene is a realistic assumption which allows predicting moving segments and their underlying motion more accurately. The overall problem however is highly non-convex. We resort to a alternating minimization technique in which we fix the motion and solve for the segments and vice versa. Finally we adapt the number of labels during the optimization using filtering and merging. Quantitative and qualitative experiments on synthetic and real world examples confirm that our method outperforms state-of-the-art scene flow methods.

Co-occurrence Priors for Continuous Multi-labeling [10]

Modern segmentation algorithms typically minimize a cost function consisting of a unary term and a regularization term. These methods promote the smoothness of an object or enforce a certain shape. Although these priors improve the solution, they are agnostic to object statistics. Hence, they do not care about objects unlikely to co-occurre in a scene. Recognizing odd object combinations comes natural to humans since we learn from a child's age to expect certain things to appear together or not. Researchers have been looking into ways to incorporate this knowledge into multi-label segmentation algorithms. The first approaches were focused on a simple semantic prior, namely the minimal description length prior [50, 106, 204, 211]. This prior reflects the philosophy of Occam's razor by encouraging parsimonious solutions, i.e. solutions with fewer labels, over more complex encodings of a scene. A more involved strategy is to exploit label co-occurrence statistics to encourage certain labels to occur jointly. This more fine grained approach of filtering out wrong labels have been proven to be very effective in the discrete setting [104]. However, being dependent on a discrete lattice, MRF approaches are known to exhibit metrication errors [95]. Additionally, extending discrete multi-labeling approaches to 3D as in [81] is not trivial because graph-cut based algorithms do not scale well with increasing dimensions. In this paper, we propose incorporating co-occurrence priors into continuous multi-label optimization. Our algorithm does not depend on the underlying grid and allows for local and non-local regularizations. The proposed formulation consists of a single convex optimization problem and can be solved in parallel using modern GPU's. This allows for fast and high quality solutions which are independent of initialization. Finally, our elegant formulation allows for an efficient extension to 3D multi-labelling.

Reprint Denied
The reprint of this publication was rejected on open-access platforms. The publication can be found at https://link.springer.com. The details are provided below.

Publication: A Co-occurrence Prior for Continuous Multi-label Optimization. In: *Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR)*. Springer Berlin Heidelberg, 2013. DOI:10.1007/978-3-642-40395-8_16

Authors:
Mohamed Souiai[1] mohamed.souiai@tum.de
Claudia Nieuwenhuis[2] cnieuwe@berkeley.edu
Evgeny Strekalovskiy[1] strekalovskiy@in.tum.de
Daniel Cremers[1] cremers@tum.de

[1]Technische Universität München, Germany
[2]ICSI, UC Berkeley, Berkeley, USA

Chapter 9

Convex Optimization for Scene Understanding

Authors:	Mohamed Souiai[1]	mohamed.souiai@tum.de
	Claudia Nieuwenhuis[2]	cnieuwe@berkeley.edu
	Evgeny Strekalovskiy[1]	strekalovskiy@in.tum.de
	Daniel Cremers[1]	cremers@tum.de

[1]Technische Universität München, Germany
[2]ICSI, UC Berkeley, Berkeley, USA

Status:	Published
Publication:	Mohamed Souiai, Claudia Nieuwenhuis, Evgeny Strekalovskiy, and Daniel Cremers."Convex Optimization for Scene Understanding". In: *International Conference on Computer Vision (ICCV), Workshop on Graphical Models for Scene Understanding*, IEEE, 2013. DOI:10.1109/ICCVW.2013.131

Individual contribution	Leading role in realizing the scientific project	
	Problem Definition	significantly contributed
	Literature survey	significantly contributed
	Implementation	significantly contributed
	Experimental evaluation	significantly contributed
	Preparation of the manuscript	significantly contributed

Abstract In this paper we give a convex optimization approach for scene understanding. Since segmentation, object recognition and scene labeling strongly benefit from each other we propose to solve these tasks within a single convex optimization problem. In contrast to previous approaches we do not rely on pre-processing techniques such as object detectors or superpixels. The central idea is to integrate a hierarchical label prior and a set of convex constraints into the segmentation approach, which combine the three tasks by introducing high-level scene information. Instead of learning label co-occurrences from limited benchmark training data, the hierarchical prior comes naturally with the way humans see their surroundings.

9.1 Introduction

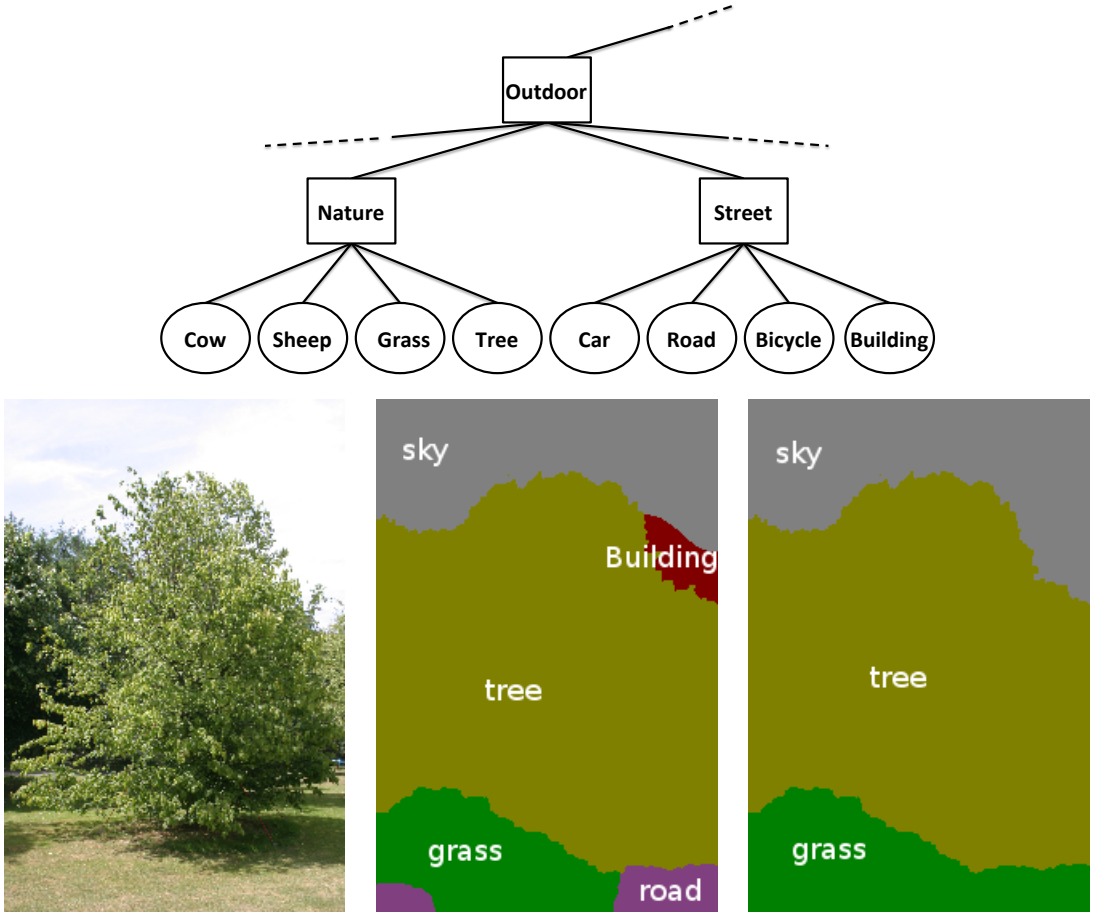
9.1.1 A Joint Approach to Scene Understanding

Scene understanding is the combination of segmentation, object recognition and scene classification. These tasks are highly interdependent. On the one hand, the most important cues for scene classification are the objects contained in the scene. On the other hand, results from scene classification help to determine the objects occurring within the scene, e.g. if we know that we are looking at a natural scene grass and sky would be likely but armchairs would be surprising. Finally, segmentation results can be improved by means of object recognition results, since typical color and shape models can be associated with the objects. Instead of solving all tasks separately or sequentially our objective is to take a holistic approach to scene understanding by solving all tasks simultaneously within a single convex optimization problem – similar to the way humans reason about the world around them. In this way, the tasks can directly influence each other. Previous joint approaches usually rely on either difficult optimization schemes or on pre-processing tasks such as superpixels or object detectors, which introduce errors and runtime limitations into the scene understanding task.

9.1.2 Related Work

The inspiration to this work predominantly draws from two lines of research, namely research on label configuration priors and research on convex relaxation techniques.

Hierarchical Semantic Prior Knowledge In human vision and understanding of the world, especially hierarchies of objects are a common concept. They can be found on a larger scene level characterizing which objects appear in a specific context, e.g. 'cars' and 'road signs' appear in 'street contexts', whereas a 'cow' and a 'sheep' usually appear in 'natural contexts' outside and not in the 'kitchen' or next to a 'computer'. But hierarchies can also be found on a small scale level describing single objects which are composed of different parts, e.g. a 'bike' consists of 'handlebars' and 'tires'. In both contexts they are characterized by specific semantic relationships among objects or object parts.



a) Input b) Potts model c) Proposed

Figure 9.1: Scene understanding consists of segmentation, object recognition and scene classification which are highly interdependent tasks. Solving these tasks within a single optimization problem such that all tasks can influence each other improves results for scene understanding. The scene in a) is classified as 'nature scene' which prevents incorrect labels such as 'building' or 'road'.

Therefore, the integration of context-related hierarchical information on the scene level is of importance to obtain highly accurate results.

The most closely related *hierarchical prior* is [50], where a fusion algorithm is proposed which computes labelings for each label group in the tree separately and fuses the results. This approach is iterative and limited to a single tree level, even though natural hierarchies consist of many levels. In addition, the algorithm exhibits optimality bounds depending on the cardinality of the label subgroups and the associated cost in each scene. This is due to the fact that with arbitrary label costs, α-expansion's bound is arbitrarily bad. For more details see [50]. In this paper we propose a non-iterative approach for trees with arbitrarily many levels and computable (in practice very tight) optimality bounds.

A special case of such hierarchical priors are *minimum description length (MDL) priors* [106, 204, 211] (with a single tree level and each class corresponding to a separate leaf with

fixed MDL cost). Such priors result in a higher penalty the more different labels occur in the image regardless of the corresponding objects.

A closely related prior is the *co-occurrence prior* [104, 10], which penalizes object sets occurring together in the same scene. The main difference to hierarchical priors is that hierarchical priors invoke a category penalty as soon as a single label of that category occurs in the scene, but they do not differentiate between labels within the same category. In contrast, co-occurrence penalties are only invoked if all labels of the specific label set occur. In addition, hierarchical priors are based on a human understanding of the world and are less complex to compute, since penalties only exist between subsequent tree levels. In contrast, co-occurrence priors are learned from limited training data and thus do not necessarily reflect general or semantically meaningful relations, but rather the label frequencies of the training set. Besides a separate penalty needs to be computed for each subset of labels (the power set), which is extremely involved and usually requires approximation [104].

Scene Classification Scene classification denotes the task of categorizing an image with respect to the type of scene shown. Most approaches build on the combination of image feature descriptors, such as color histograms, texture or SIFT features. Based on the descriptor output learning based approaches such as Support Vector Machines or statistical approaches are applied to classify the scene based on training data [21, 113, 114]. Yet, these approaches rarely solve the segmentation, object recognition and scene classification tasks jointly.

Joint Approaches Joint approaches for segmentation, recognition and scene classification were given recently in [104] and [202]. Both approaches rely on the result of sophisticated object detectors in order to infer solutions for the joint task. Thus, the quality of the results always depends on the quality of the object detectors. In addition, the inference problem solved in [104] is rather complex and does not involve the actual scene classification task.

In this paper we solve the joint task based on convex optimization techniques without requiring any preprocessing such as object detection or superpixel computation. In this way, the quality of the solutions as well as the runtime directly depend on the proposed algorithm instead of prior processing steps.

Convex Optimization To tackle the highly complex task of joint recognition and scene classification we will rely on powerful techniques from convex optimization. In general, this scene understanding task can be formulated as a multi-label problem. Two popular paradigms exist for solving such energy optimization problems: discrete Markov Random Field (MRF) based approaches and continuous optimization approaches. In [137] Nieuwenhuis et al. showed that for multi-label problems continuous approaches can be parallelized and implemented more efficiently. In addition, they do not suffer from grid bias and - in case of a convex relaxation - are independent of the initialization. There are a number of recent advances on convex relaxation techniques for spatially continuous multi-label optimization. These include relaxations for the continuous Potts model [38, 107, 206], for

the non-local continuous Potts model [199], for MDL priors [204], and for vector-valued labeling problems [70, 174]. In this paper we will give a convex relaxation of the scene understanding task.

9.1.3 Contribution

Our main contributions are the following:

- Instead of solving a sequence of optimization problems we introduce the hierarchical segmentation prior within a *single* convex optimization problem.
- The performance of our algorithm neither depends on the label subset cost nor on the cardinality of the label subsets and can therefore handle an arbitrary number of labels.
- Our formulation is more general than the class of hierarchical priors in [50] in the sense that we are able to assign arbitrary costs for label configurations arising from different categories. We also introduce a variant of the hierarchical prior where we even assign infinite costs to certain label configurations.

9.2 Convex Multi-label Segmentation

Given a discrete label space $\mathcal{L} = \{1, ..., n\}$ with $n \geq 1$, the multi-labeling problem can be stated as a minimal partition problem. The image domain $\Omega \subset \mathbb{R}^2$ is to be segmented into n pairwise disjoint regions Ω_i which are encoded by the label indicator function $u \in BV(\Omega, \{0, 1\})^n$

$$u_i(x) = \begin{cases} 1 & \text{if } x \in \Omega_i, \\ 0 & \text{otherwise.} \end{cases} \tag{9.1}$$

Here BV denotes the space of functions with bounded total variation, which allows for discontinuities [16]. To ensure that each pixel is assigned to exactly one region, the simplex constraint is imposed on u:

$$\sum_{i=1}^{n} u_i(x) = 1 \quad \forall x \in \Omega \tag{9.2}$$

To find a solution to the minimal partition problem we minimize the following energy:

$$E(u) = E_D(u) + E_S(u) + E_H(u). \tag{9.3}$$

The data term

$$E_D(u) = \sum_{i=1}^{n} \int_{\Omega} u_i(x) \varrho_i(x) dx. \tag{9.4}$$

where $\varrho_i(x)$ is the local cost of assigning label i to pixel x, measures how well the segmentation complies with a given appearance model for each label. The regularizer:

$$E_S(u) = \frac{1}{2}\sum_{i=1}^{n}\int_\Omega |Du_i(x)| \tag{9.5}$$

ensures spatial coherence of the label assignment, and is chosen as the Potts model which penalizes the boundary lengths. The term $E_H(u)$ is the hierarchical scene understanding energy which will be the focus of this paper.

Label Occurrence Functions In order to devise the hierarchical prior it is necessary to model the occurrences of specific labels in the image. Let $\mathcal{U} = \{u \in BV(\Omega, \{0,1\})^n \mid \sum_i u_i(x) = 1 \ \forall x \in \Omega\}$ denote the set of all possible segmentations over the image domain Ω. Then the function $l : \mathcal{U} \to \{0,1\}^n$ indicates for each label $i \in \{1, \ldots, n\}$ whether it occurs in a given segmentation:

$$l_i(u) = \begin{cases} 1 & \text{if } \exists x \in \Omega : \ u_i(x) = 1, \\ 0 & \text{otherwise.} \end{cases} \tag{9.6}$$

This can also be written as [204]

$$l_i(u) = \max_{x\in\Omega} u_i(x) \quad \forall i \in \mathcal{L} \tag{9.7}$$

9.3 The Hierarchical Prior

Hierarchical priors penalize the co-occurrence of labels from different scene contexts, e.g. a 'cow' ('outdoor' context) and a 'fridge' ('indoor' context). To this end, the set of object labels is organized in a tree structure where the leaves correspond to objects and the inner nodes to object categories $\mathcal{S}$ with $k := |\mathcal{S}|$, see Figures 9.1 and 9.2.

Let $\pi : \mathcal{S} \to \mathcal{L}$ maps each category to the set of object labels it contains in *all* of its subtrees, e.g. $\pi(L_6) = \{l_5, l_6, l_7, l_8\}$ in Figure 9.2. Let furthermore

$$L : \mathcal{U} \to \{0,1\}^k, L_i(u) = \max_{j\in\pi(i)} l_j(u) \tag{9.8}$$

denote the indicator function for the k categories in the inner tree nodes, i.e. $L_i(u)$ indicates if any label in any subtree of category L_i is present in the scene. These nodes are organized in arbitrarily many levels, e.g. 'outdoor' contains the subcategories 'nature' and 'street' (see Figure 9.1). Note that labels can be shared by several categories by adding them one level above all the subtrees that should share them. See for example the labels 'sky' and 'grass' in Figure 9.3, which can appear in 'nature', 'street' and 'water' scenes.

For each single category function L_i we define a specific cost $C_{L_i} \geq 0$ which is added to the energy if any of the objects in any subtree of the category L_i appears in the segmentation.

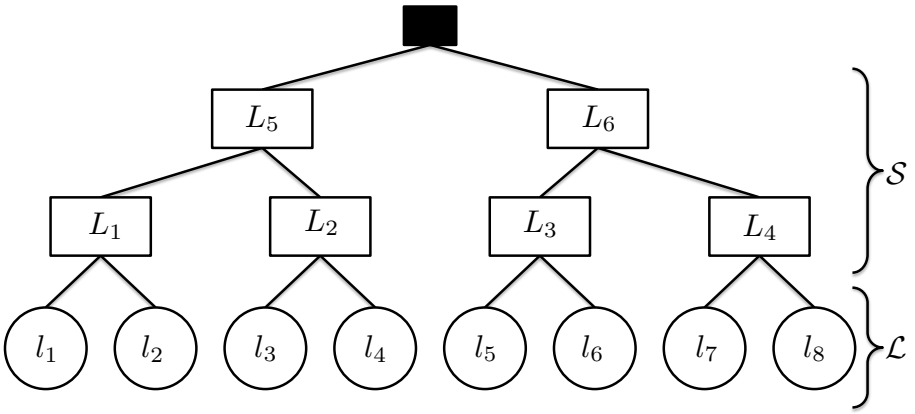


Figure 9.2: An example label hierarchy with object labels $\mathcal{L}$ in the leaves and scene labels $\mathcal{S}$ in the inner nodes.

Hence, if the label 'bicycle' appears the costs for the categories 'street' and 'outdoor' are invoked.

Then we can define the hierarchical energy as

$$E_H(u) := \sum_{i \in \mathcal{S}} C_{L_i} L_i(u) \tag{9.9}$$

with each L_i given by (9.8). Thus, for each label occurring in the segmentation the energy is increased by the costs C_{L_i} for all categories L_i the label belongs to. In this way, we can introduce statistical information on the likelihood of different scenes. Here, conditional likelihoods instead of absolute ones are of interest, i.e. the probability of a scene given its direct parent in the tree. For the optimization, we use (9.8) and (9.7) to write $L_i = \max_{j \in \pi(i), x \in \Omega} u_j(x)$ in terms of u. For efficient minimization we decouple max-terms by means of the dualization of the max function [204].

Scene Uniqueness Constraints The hierarchical prior introduces costs depending on the likelihood of each scene. In this way, it discourages labels from different scenes but nevertheless allows for mixed solutions. For scene classification, however, one would expect a hard decision for a single scene label. In order to obtain a unique scene label and to improve the segmentation at the same time, we propose to introduce a scene uniqueness prior. This prior imposes the constraint that all labels occurring in the final segmentation belong to the same category, i.e. they share the same path from the lowest category level to the root node of the tree. Let $P : \mathcal{S} \to \mathcal{S}$ map all categories to their direct category child nodes, i.e. 'outdoor' is mapped to 'street' and 'nature', and the lowest categories such as 'water' are mapped to the empty set. Then we impose the following constraints

$$\sum_{j \in P(i)} L_j(u) = L_i(u) \quad \forall i \in \mathcal{S}. \tag{9.10}$$

This constraint set ensures that the sum of all category functions at each tree level equals the parent category function. By setting the root node indicator function $L_r(u) = 1$ we enforce a unique scene classification result. If a subcategory function is zero then no label from its subtree can occur in the segmentation result. If two labels in different subtrees are active then the scene uniqueness constraints force one of them to zero. These constraints are linear and can be easily implemented by means of Lagrange multipliers. They can be applied in addition to the hierarchical prior or alone. The energy E_H then reads as:

$$E_H(u) := \sum_{i \in \mathcal{S}} C_{L_i} L_i(u) \quad \text{s.t.} \sum_{j \in P(i)} L_j(u) = L_i(u), \tag{9.11}$$

$$L_i(u) = \max_{j \in \pi(i)} l_j(u), \;\; l_i(u) = \max_{x \in \Omega} u_i(x), \quad \forall i \in \mathcal{L}. \tag{9.12}$$

In addition to u, the overall energy (9.3) is then also optimized over the indicator functions l_i and L_i as new variables. Since the max-constraints (9.12) are not convex, we replace them by the relaxations

$$l_j(u) \leq L_i(u) \quad \forall i \in \mathcal{S},\, j \in \pi(i), \tag{9.13}$$

$$u_i(x) \leq l_i(u) \quad \forall i \in \mathcal{L},\, x \in \Omega. \tag{9.14}$$

They can be implemented with Lagrange multipliers, e.g. by adding the terms $\sup_{a_i(x) \geq 0} \int_\Omega a_i(x)(u_i(x) - l_i(u))\, dx$ to the energy (9.3) and optimizing also over a.

9.4 Implementation

In order for the domain of optimization to be a convex set, we relax the binary constraint $u_i(x) \in \{0, 1\}$ to the convex one $u_i(x) \in [0, 1]$. To minimize the overall energy (9.3) we use the primal-dual algorithm [39], which is essentially a gradient descent in the primal variables and a gradient ascent in the dual variables with a subsequent application of the proximity operators. For the time steps we used the recent preconditioning [146].

9.5 Experiments

We will now show results for the joint task of segmentation, object recognition and scene classification. To this end, we have selected a set of 15 semantic labels and scene types (out of 21), which could naturally be grouped in a tree hierarchy, see Figure 9.3. Hence, we define the following label set

$$\begin{aligned}\mathcal{L} := \{&\text{Grass,Car, Bird, Building, Sky, Water,Cow,Sheep,Boat,} \\ &\text{Chair, Tree, Sign, Road, Book,Sky}\}\end{aligned}$$

together with the object categories

$$\mathcal{S} := \{\text{Indoor, Outdoor, Nature, Street}\}.$$

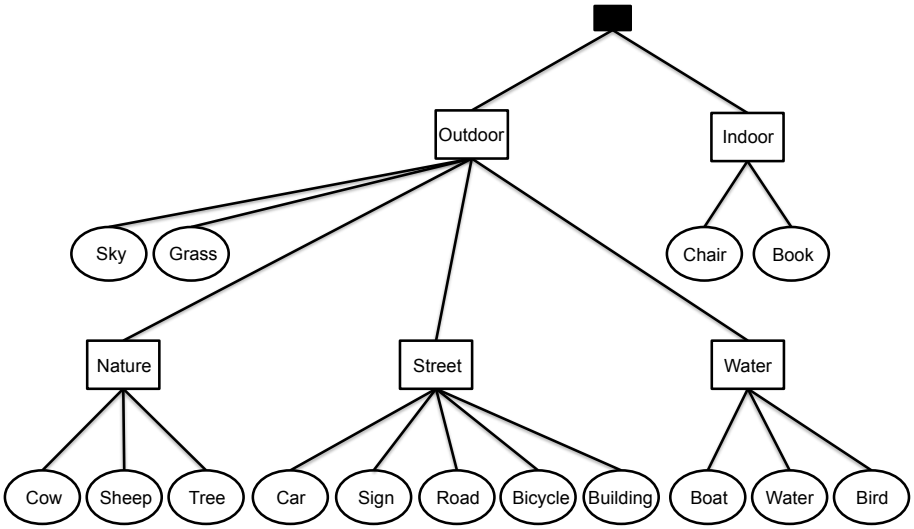


Figure 9.3: **Hierarchical prior** for MSRC benchmark used for joint segmentation, recognition and scene classification.

	Global	Per class	Building	Grass	Tree	Cow	Sheep	Sky	Water	Car	Bicycle	Sign	Bird	Book	Chair	Road	Boat
Potts	87.72	79.09	79	97	**89**	70	81	**97**	74	**95**	81	85	**70**	82	**96**	65	32
Co-occurrence [104]	**89.97**	81.76	**88**	**99**	86	62	86	92	**94**	94	82	**89**	62	**88**	84	71	34
Hierarchical Prior	89.53	**81.83**	82	97	**89**	**82**	**91**	90	89	**95**	**90**	88	64	87	56	**79**	**41**

Figure 9.4: Average accuracies over all images (global) and average per class for the pure Potts model, our approach and the co-occurrence results by Ladicky et al. [104]. The scores for each label are defined as $\frac{True\ Positives \cdot 100}{True\ Positives + False\ Negatives}$.

For testing we use the subset of 68 images from the MSRC benchmark, which contains only labels within our hierarchy. We minimize the energy in (9.3) using the appearance term in [104] as data term E_D, the standard Potts model as smoothness term $E_S(u)$ and the hierarchical energy E_H together with scene uniqueness constraints as formulated in (9.11). Qualitative results comparing the proposed approach to the results based only on the Potts model (i.e. $E_H = 0$) and to the co-occurrence priors by Ladicky et al. [104] are shown in Figure 9.5. Several of these images show strong improvements compared to the Potts and the co-occurrence prior, e.g. the 'book' label disappears from the sign image, the reflection of the tree is correctly classified as 'water' and the 'sheep' is no longer confused with the label 'road' due to color similarities. Figure 9.4 shows the average accuracy on the mini-benchmark. The comparison to the co-occurrence prior shows that the differences are only marginal on average. Yet there is no scene classification involved in the co-occurrence prior, and as argued in the introduction learning of the prior is much more involved and prone to specialization on the specific database. In contrast, the hierarchy structure was modeled by hand based on human reasoning.

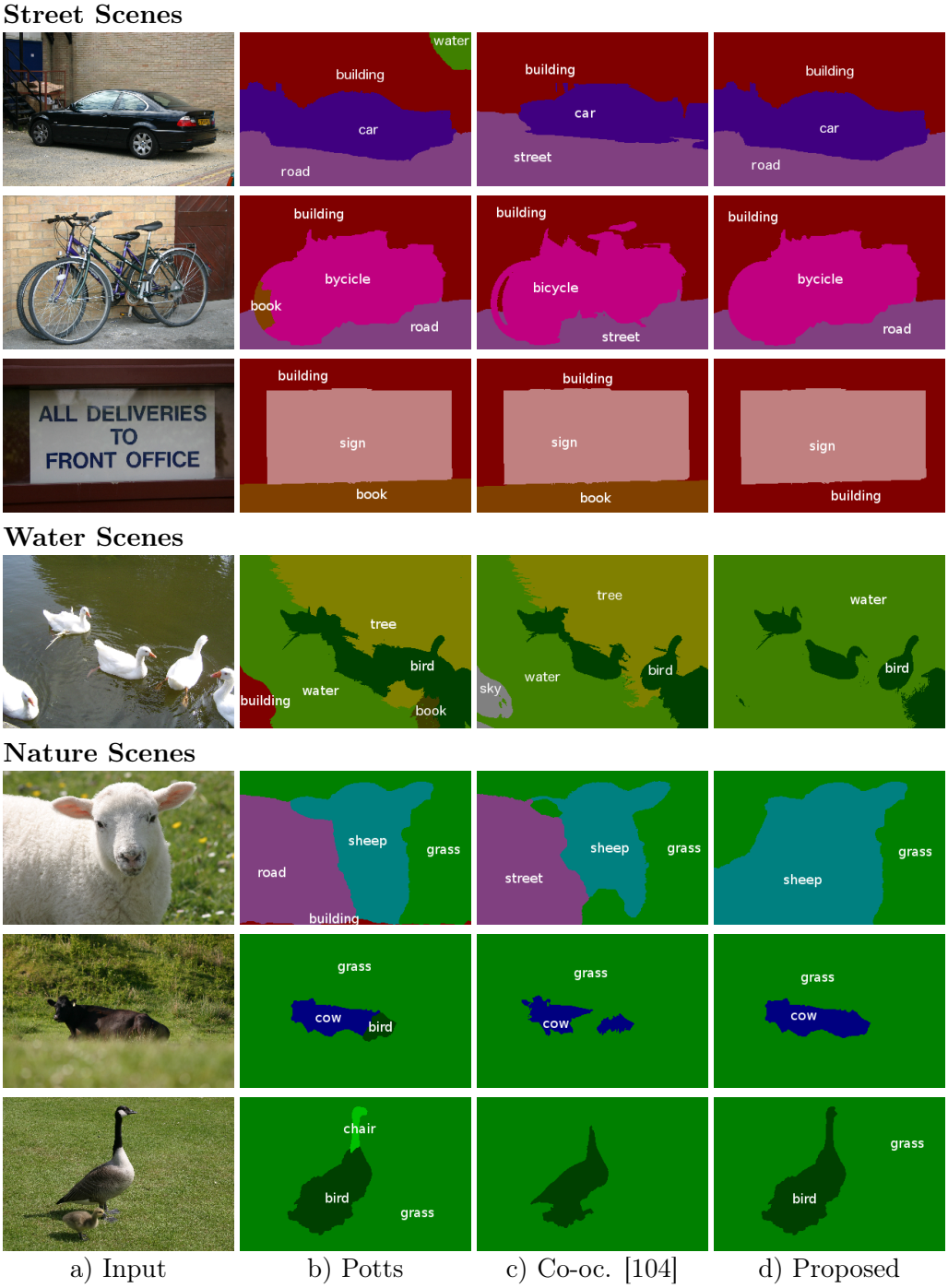


Figure 9.5: **MSRC benchmark** results for joint segmentation, recognition and scene classification using the hierarchy in Figure 9.3.

9.6 Conclusion

In this paper we proposed a joint approach for segmentation, object recognition and scene understanding, which is formulated within a single multi-label variational optimization approach. In contrast to previous approaches we do not rely on the computation of superpixels or object detector outputs or build several stages in the optimization process. We gave a convex relaxation of the approach yielding unique solutions to the scene classification task independent of the initialization of the algorithm. The results on the MSRC benchmark show that for several images we were able to strongly improve the labeling task achieving classification results sligtly above the highly specialized co-occurrence prior by Ladicky et al. [104].

Chapter 10

Entropy Minimization for Multi-Label Optimization

Authors:	Mohamed Souiai[1]	mohamed.souiai@tum.de
	Martin Oswald[2]	martin.oswald@inf.ethz.ch
	Youngwook Kee[4]	yok2013@med.cornell.edu
	Junmo Kim[3]	junmo.kim@kaist.ac.kr
	Marc Pollefeys[2]	marc.pollefeys@inf.ethz.ch
	Daniel Cremers[3]	cremers@tum.de

[1]Technische Universität München, Germany
[2]ETH Zürich, Switzerland
[3]KAIST, South Korea
[4]Cornell University

Status:	Published
Publication:	Mohamed Souiai, Martin R. Oswald, Youngwook Kee, Junmo Kim, Marc Pollefeys, and Daniel Cremers. "Entropy minimization for convex relaxation approaches". In: *International Conference on Computer Vision (ICCV)*. IEEE, 2015. DOI: 10.1109/ICCV.2015.207

Individual contribution	Leading role in realizing the scientific project	
	Problem definition	significantly contributed
	Literature survey	significantly contributed
	Implementation	significantly contributed
	Experimental evaluation	significantly contributed
	Preparation of the manuscript	significantly contributed

Abstract Despite their enormous success in solving hard combinatorial problems, convex relaxation approaches often suffer from the fact that the computed solutions are far from binary and that subsequent heuristic binarization may substantially degrade the quality of computed solutions. In this paper, we propose a novel relaxation technique which incorporates the entropy of the objective variable as a measure of relaxation tightness. We show both theoretically and experimentally that augmenting the objective function with an entropy term gives rise to more binary solutions and consequently solutions with a substantially lower optimality gap. We use difference of convex function (DC) programming as an efficient and provably convergent solver for the arising convex-concave minimization problem. We evaluate this approach on three prominent non-convex computer vision challenges: multi-label inpainting, image segmentation and spatio-temporal multi-view reconstruction. These experiments show that our approach consistently yields better solutions with respect to the original integral optimization problem.

10.1 Introduction

Numerous problems in vision – including two-region and multi-region image segmentation, stereo- and multi-view reconstruction or optical flow estimation – can be cast as variational multi-labeling problems of the form

$$\mathbf{u}^* = \arg\min_{\mathbf{u}} E(\mathbf{u}), \tag{10.1}$$

with a labeling function $\mathbf{u} : \Omega \to \Gamma$ from a domain Ω to a label space Γ. The domain and the label space can be both discrete or continuous. Numerous works have proposed methods for efficiently computing solutions to the above integer problem. A common approach [151] is to rephrase the above integer problem as a binary labeling problem with an indicator variable $u : \Omega \to \{0,1\}^{|\Gamma|}$:

$$u^*_{\text{bin}} = \arg\min_{u:\Omega\to\{0,1\}^{|\Gamma|}} E(u) \ , \tag{10.2}$$

possibly subject to additional constraints. While this binary formulation often leads to a convex energy E, the binary constraint makes the optimization domain non-convex and generally yields a hard combinatorial problem.

Relaxation. The central idea of many convex relaxation techniques is to drop the integrality constraint and consider the relaxed convex problem:

$$u^*_{\text{rel}} = \arg\min_{u:\Omega\to[0,1]^{|\Gamma|}} E(u). \tag{10.3}$$

Solving this problem is usually much easier, but its solution u^*_{rel} can be non-binary and consequently a rounding scheme needs to be applied to obtain a binary solution.

Rounding. In general, the rounded solution can be very different from the globally optimal solution u^*_{bin} and may not even be a local minimum of the binary objective (10.2). The simplest rounding scheme is to select the label with the highest likelihood. We will detail the rounding schemes later when looking at certain problem instances of (10.3).

For certain 2-label problems, a thresholding theorem [138] may assure provably optimal binary solutions upon simple thresholding of the relaxed solution u^*_{rel}. For more general multi-label problems the computed solutions are often far from binary and rounding may drastically increase the energy and the corresponding a-posteriori optimality bound (the energetic difference between rounded and relaxed solution). Moreover, a number of works incorporate additional constraints on the solution in order to constraint the volume/area [155, 186], or higher order moments [96] of the resulting segmentation, or to enforce size proportions of respective segments [136]. While these constraints are meaningful for the binary problem (10.2), they often change their physical meaning in the relaxed setting (10.3).

Novel relaxation scheme. In order to cope with the above mentioned problems we propose to augment problem (10.1) with an additional term which promotes the integrality of the solution during optimization and thereby leads to better solutions with substantially smaller optimality gaps. The key idea is to control the integrality of the objective variable by means of Shannon's entropy. Entropy minimization has been used in several computer vision applications including shadow removal [63] and image segmentation [180] where it is used as a general color consistency criterion for separating histograms. To the best of our knowledge this work is the first to apply entropy minimization on the labeling function u. The augmented relaxed problem composes of the original labeling problem and the additional entropy term H weighted by $\theta \in \mathbb{R}_{\geq 0}$ and reads as follows:

$$u^* = \underset{u:\Omega \to [0,1]^{|\Gamma|}}{\arg\min} \; E(u) + \theta H(u). \tag{10.4}$$

This generalizes the convex relaxation problem (10.3), obtained by setting $\theta = 0$. Likewise, by taking $\theta \to \infty$ we also recover (10.2) since the entropy term becomes an integral constraint. Unfortunately problem (10.4) is not convex anymore for $\theta > 0$ and obtaining the global minimum of its relaxed version is generally not possible anymore. Luckily, problem (10.4) exhibits a special structure: it decomposes into a convex function $E(\cdot)$ and a concave function $\theta H(\cdot)$ being equivalent to a difference of convex functions (DC). This makes formulation (10.4) amenable to the DC programming approach [182] which guarantees to find a stationary point of the objective. Although there is no guarantee for finding a globally optimal solution, we show that the obtained non-rounded results are more binary than the results of state-of-the-art relaxation methods. In this paper, we focus on spatially continuous variational approaches as they are easily parallelizable and do not suffer from metrication errors in contrast to their discrete counterparts [95]. Though, our approach might also be applicable in a discrete setting.

10.1.1 Contributions

Our contributions can be summarized as follows:

- We propose to use the entropy of the objective variable to empirically measure the tightness of a relaxed solution. We then enhance integrality of solutions by jointly optimizing the objective function and the entropy.

- We propose to use a provably convergent algorithm for solving the arising convex-concave problem which combines DC programming and a state-of-the-art first order solver.
- We show theoretically and experimentally that our method provides solutions which are more integral and exhibit a tighter energy bound than state-of-the-art convex relaxation methods. Our approach thus promotes simple rounding schemes.
- The proposed entropy augmentation does not change the algorithm complexity. In all experiments we observed an improved convergence behavior and runtime speed-ups up to a factor of two.
- We demonstrate the effectiveness of our approach on several computer vision applications including multi-label image inpainting, image segmentation and spatio-temporal multi-view reconstruction.

10.2 Shannon's Entropy

Originally formulated for discrete random variables Shannon's information entropy is a measure of uncertainty. Suppose that $u : \Omega \subset \mathbb{R}^d \to [0,1]^{|\Gamma|}$ is a soft labeling function such that $\forall x \in \Omega : \sum_{\ell=1}^{|\Gamma|} u_\ell(x) = 1$. Interpreting $u(x)$ as a probability distribution on each $x \in \Omega$ we can apply information entropy on a labeling problem by directly imposing it on the indicator variable u. The total entropy of a labeling function can be written as follows:

$$H(u) = \int_\Omega -\sum_{\ell=1}^{|\Gamma|} u_\ell \log u_\ell \; dx \;\; , \tag{10.5}$$

which is an integral of a point-wise concave entropy measure in each $x \in \Omega$. For brevity we will refer to the total entropy as entropy for the rest of the paper.

10.3 Solving the Convex-Concave Program

In the field of variational convex relaxation approaches, non-convex optimization has gained tremendous popularity in recent years. However most of these works focus on realizing non-convex regularizers [129, 140]. We make use of DC programming, dating back to a seminal work by Tao *et al.* [182] which generalizes subgradient algorithms for convex maximization. The principle of minimizing the difference of convex functions heavily relies on concepts from convex optimization and especially DC duality [185]. Closely related to DC programming is the so called convex-concave procedure (CCCP) described later in [205] though it assumes differentiability of the objective function. In [92] DC programming is applied to a QP relaxation of MAP inference in order to cope with the non-convex objective.

10.3.1 DC Programming

DC programming deals with solving a non-convex problem of the following form:

$$\inf_u \{g(u) - h(u)\} \ , \tag{10.6}$$

where $g(u)$ and $h(u)$ are convex functions. In order to solve (10.6) we make use of a simplified form of the DC algorithm [181]. Based on DC duality and the KKT conditions for DC programs, the algorithm generates the following sequences $v^k \in \partial h(u^k)$ and $u^{k+1} \in \partial g^*(v^k)$ which guarantee to converge to a critical point. The overall DC algorithm is illustrated in algorithm 5, where g^* denotes the Legendre-Fenchel conjugate of g.

Algorithm 5 The DC Algorithm

Initialize $u^0 \in \text{dom } g$.
while not converged **do**
  Choose $v^k \in \partial h(u^k)$
  Choose $u^{k+1} \in \partial g^*(v^k)$
  $k \leftarrow k + 1$
end while

10.3.2 Solving monotone inclusions

Note that since $u^{k+1} \in \partial g^*(v^k) \Leftrightarrow v^k \in \partial g(u^{k+1}) \Leftrightarrow 0 \in \partial g(u^{k+1}) - v^k$ and since g a convex, we solve a monotone inclusion problem in each iteration. Hence choosing $u^{k+1} \in \partial g^*(v^k)$ in algorithm 5 amounts to solving the following convex optimization problem:

$$u^{k+1} = \arg\min_u g(u) - \langle v^k, u \rangle \tag{10.7}$$

In order to be able to solve problem (10.4) using DC programming we make the following identifications:

$$g(u) = E(u) + \delta_{U_{\text{rel}}}(u) \qquad h(u) = \theta H(u) \ , \tag{10.8}$$

where $\delta_{U_{\text{rel}}}(u)$ denotes the characteristic function:

$$\delta_{U_{rel}}(u) = \begin{cases} 0 & \text{if } u(x) \in [0,1] \ \ \forall x \in \Omega, \\ \infty & \text{otherwise.} \end{cases} \tag{10.9}$$

Most variational problems $g(u)$ are large scale and non-smooth and therefore not easily solvable using standard solvers. To this end, we use the state-of-the-art primal-dual algorithm of Pock *et al.* [39] which solves a saddle point formulation of problem (10.7) (further details in the supplementary material). Using Algorithm 5 we can solve any problem of the form (10.4). For binary tomographic reconstruction, [196] proposed a similar DC programming framework for implicit rounding. In contrast to the entropy term which is well-grounded on information theory and naturally generalizes to higher dimensions, they use a negative quadratic term as a heuristic to impose integrality of the solution.

10.4 Experiments

We consider three problem instances which exhibit non-tight relaxations. For all experiments we initialized function u in our algorithm with a zero or random function while choosing the entropy parameter as $\theta \in [0.01, 0.5]$.

10.4.1 Multi-label Image Segmentation

In variational multi-label image segmentation one assumes a continuous domain Ω and a discrete label space Γ with $|\Gamma| \geq 2$. The image domain $\Omega \subset \mathbb{R}^2$ is to be segmented into $|\Gamma|$ pairwise disjoint regions Ω_ℓ which are encoded by the label indicator function $u : \Omega \to \{0,1\}^{|\Gamma|}$, $u_\ell(x) = \mathbf{1}_{x \in \Omega_l}$. The overall problem can be stated as a minimal partition problem. To find a solution to such a problem augmented with an entropy term we solve the following optimization problem:

$$\begin{aligned} \min_u & \sum_{\ell=1}^{|\Gamma|} \int_\Omega \left[\lambda u_\ell(x) \varrho_\ell(x) + \frac{1}{2} |Du_\ell| \right] dx + \theta H(u) \\ \text{s.t.} & \sum_{\ell=1}^{|\Gamma|} u_\ell(x) = 1, \quad u(x) \geq 0 \qquad \forall x \in \Omega \end{aligned} \tag{10.10}$$

The data fidelity term $\varrho_l(x) : \Omega \to \mathbb{R}$ assigns a color-based pixel-wise cost to each pixel x for belonging to region ℓ. Expression Du in the smoothness term denotes the distributional

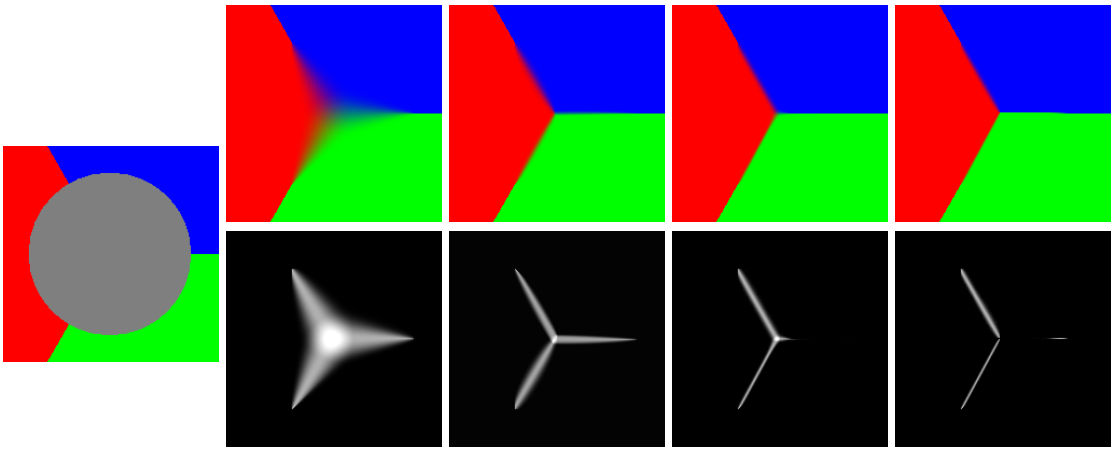


Figure 10.1: Inpainting using different relaxations of the Potts model with ($\theta = 0.02$) and without ($\theta = 0$) entropy. The visualized entropy shows that the joint minimization of the entropy yields more binary solutions.

derivative of u throughout the paper. It encourages regularity of the obtained partitions in the solution by minimizing its boundary length and is chosen as proposed by Zach *et al.* [206]. Although the relaxation of Chambolle, Cremers and Pock (CCP) [38] of the

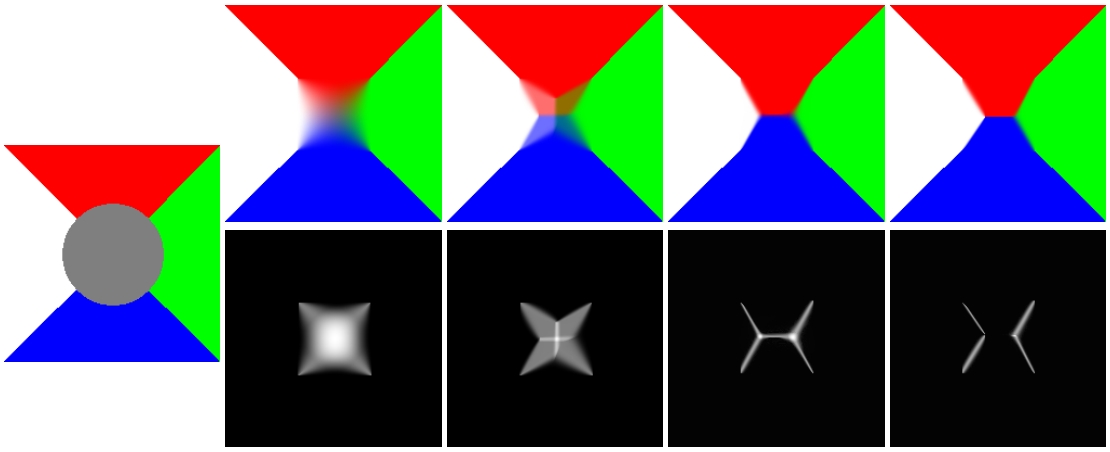


Figure 10.2: Inpainting in the case of 4 regions. In case of non uniqueness our approach (with $\theta = 0.02$) picks an almost integral solution with lower energy (see table 10.1).

boundary length is tighter compared to Zach *et al.* [206], its complexity grows quadratically with the number of labels, making it impractical for large scale problems with many labels. Note that the energy functional (10.10) is a convex-concave program which can be solved by applying the DC algorithm 5. Figure 10.1 illustrates that despite the less tight relaxation (10.10) we obtain a drastic improvement of the solution by augmenting the optimization problem by an entropy term. We observe that for the inner optimization problem in Algorithm 5, only a few iterations (1-5) are necessary for convergence. As a result, we obtain tight solutions even with the simple relaxation of Zach *et al.* [206] without the need to drastically increase the runtime by using a tighter CCP relaxation.

Rounding. To compute a binary solution from the relaxed one, we select the most likely label point-wise, i.e. $\forall x : u_{\text{bin}}(x) = \hat{\mathbf{e}}_j$ with $j = \min\left\{ \arg\max_\ell (u^*_{\text{rel}}(x))_\ell \right\}$ and $\hat{\mathbf{e}}_j$ being the j-th unit vector in the space $\{0,1\}^{|\Gamma|}$. Note that Lellmann *et al.* [111] proposed a probabilistic rounding scheme which provides an a-priori bound on the energy. However it is slower and yields slightly degraded results in practice (see [110] for a comparison on the triple-junction problem).

Non-Uniqueness of the Solution

The energy (10.10) without $H(u)$ is not strictly convex and thus admits several binary solutions. If the relaxation is not tight then convex combinations of distinct binary solutions may get assigned lower energies which in turn promotes non-binary solutions. This is especially visible for larger numbers of labels. Figure 10.2 shows that already with 4 labels the CCP relaxation tends to produce a convex combination of binary solutions which is a valid minimizer in the case of the relaxed problem as we show in Proposition 1.

Proposition 1 (Convex combinations of binary solutions have lower energies)**.** *The convex*

combination $u_\alpha = \alpha u_1^* + (1-\alpha)u_2^*$ *of two binary solutions* $u_1^*, u_2^* \colon \Omega \to \{0,1\}^{|\Gamma|}$ *of problem* (10.2) *has lower or equal energy (Eq.* (10.3)*) than the energies of the binary solutions, i.e.* $E(u_\alpha) \leq E(u_1^*)$.

Proof. By convexity of the objective $E(u)$ we have:
$E(\alpha u_1^* + (1-\alpha)u_2^*) \leq \alpha E(u_1^*) + (1-\alpha)E(u_2^*)$
and since $E(u_1^*) = E(u_2^*)$ we prove the claim. The generalization for more than two solutions is straightfoward. □

We observe that adding an entropy term tends to constrain the solution space to more integral solutions, hence our approach helps picking a solution which is more binary and hence tighter than the convex combination. In addition to the entropy, we measure the tightness of the solution by also evaluating the established posterior optimality bound [137] $G(u_{\text{bin}}, u_{\text{rel}}^*) = \frac{E(u_{\text{bin}}) - E(u_{\text{rel}}^*)}{E(u_{\text{rel}}^*)}$, where u_{rel}^* is the solution of the relaxed problem and u_{bin} the corresponding rounded solution. The exact runtimes and comparisons of different relaxations combined with entropy are presented in Table 10.1. Note that in addition to obtaining tighter solutions our algorithm outperforms the original formulation even in the runtime. The following Proposition 2 shows that additionally minimizing the entropy promotes binary solutions over convex combinations.

Proposition 2 (Energy (10.4) favors binary solutions over convex combinations)**.** *Lets denote the objective* (10.4) *by* $E_H(u) = E(u) + \theta H(u)$ *and let* $u_\alpha = \sum_i \alpha_i u_i^*$ *be a convex combination of different optimal binary labelings* $u_i^* \colon \Omega \to \{0,1\}^{|\Gamma|}$ *with normalized weights* $\alpha_i \in [0,1]$, $\sum_i \alpha_i = 1$. *Then, for sufficiently large* θ*, binary solutions* u_i^* *have a strictly lower entropy-augmented energy* E_H *than convex combinations of binary solutions* u_α*, that is,* $E_H(u_\alpha) > E_H(u_i^*)$.

Proof. Using the definition of E_H, the convexity of $E(u)$ and the property that $H(u)$ vanishes for binary u, we derive the following equalities and inequalities for any binary solution u_i^*:

$$E_H(u_\alpha) - E_H(u_i^*) = \underbrace{E(u_\alpha) - E(u_i^*)}_{\leq 0} + \theta \underbrace{\left(H(u_\alpha) - H(u_i^*)\right)}_{\geq 0} \tag{10.11}$$

By choosing $\theta > \frac{E(u_i^*) - E(u_\alpha)}{(H(u_\alpha) - H(u_i^*))}$ the following inequality holds:

$$E_H(u_\alpha) - E_H(u_i^*) > 0 \Leftrightarrow E_H(u_\alpha) > E_H(u_i^*) \tag{10.12}$$

□

Table 10.1 shows the corresponding energies E_{bin} and E_{rel} to the results in Figure 10.2 for the binary and relaxed solutions for the Zach and CCP relaxation respectively. While the energies obtained using the additional entropy term are higher than the relaxed energies, the binarized energies of both relaxations combined with an entropy penalization are clearly lower than the mere convex relaxations.

	Zach	CCP	Zach+Entropy	CCP+Entropy
E_{rel}	630.1	634.3	639.6	**637.7**
E_{bin}	673.9	691.6	**666.8**	668.1
$G(u_{\text{bin}}, u^*_{\text{rel}})$	0.069	0.090	**0.042**	0.047
Entropy H	2769	1739	379	**306**
Runtime [s]	153	188	**68**	125

Table 10.1: Relaxed and binary energies for different relaxations with and without entropy term, optimality gaps as well as entropy values and runtimes in seconds for the 4-region inpainting in Fig. 10.2.

10.4.2 Binary Image Segmentation with a Fixed Volume Constraint

By considering only two labels the relaxation of problem (10.3) becomes tight and optimal binary solutions can be computed via simple thresholding of the relaxed solution [138]. Unfortunately, this changes easily by adding further constraints to the optimization problem. We consider a fixed volume constraint on the solution of the segmentation problem. Volume constraints have been used with convex relaxation methods for image segmentation [134, 155], image-based modeling [143, 186] and they have also been generalized to higher order moment constraints [96]. In [134] they also addressed the problem of the non-tight relaxation, but their suggested algorithm is less general as user-provided seed points are required. Discrete approaches to this NP-hard problem have been suggested in [56, 115]. Both of them are restricted to equality constraints and the former only provides approximate solutions and its runtime is exponential in the number of labels. Our approach is more general than previous works, it provides the desired results for the fixed volume segmentation problem and it improves the convergence of a state-of-the-art solver at the same time. Thus, we consider the following minimization problem:

$$\begin{aligned} &\min_u \int_\Omega \left[g|Du| + \lambda f u \right] dx + \theta H(u) \\ &\text{s.t.} \int_\Omega u \, dx - V_t = 0 \ , \end{aligned} \tag{10.13}$$

where λ steers the smoothness of the solution by changing the impact of the data fidelity term being defined by function $f : \Omega \to \mathbb{R}$, $f(x) = (c_1 - I(x))^2 - (c_2 - I(x))^2$ in which $c_1, c_2 \in \mathbb{R}$ are gray values for foreground and background, $I : \Omega \to \mathbb{R}$ is the input image and V_t denotes the pre-defined target volume. The total variation weight is defined as $g = \exp(-|\nabla f|)$. Note that one easily adapts the approach to bound the volume with inequality constraints [96, 134], but the problems of the relaxation can be better demonstrated with an equality constraint.

Because of the volume constraint there is no free choice of thresholds to obtain a binary solution and the thresholding theorem [138], which directly relates solutions of the relaxed problem to binary one, does not apply anymore. For rounding, simple thresholding as in [56] might violate the volume constraint. The following rounding scheme from [186]

addresses both the volume constraint and the binary constraint. It even guarantees fulfillment of the volume constraint, if several non-binary $u(x)$ have identical values, i.e. no threshold exists to binarize the solution without violating the volume constraint.

Proposition 3 (Rounding scheme for fixed volume constraint [186, Prop.2])**.** *The relaxed solution of* (10.13) *can be projected to the set of binary functions in such a way that the resulting binary function preserves the target volume* V_t*.*

Proof. It suffices to order the voxels $x_i \in V$ by decreasing values $u(x_1) \geq u(x_2) \geq \ldots \geq u(x_{|V|})$. Subsequently, one sets the value of the first V_t voxels to 1 and the value of the remaining voxels to 0. □

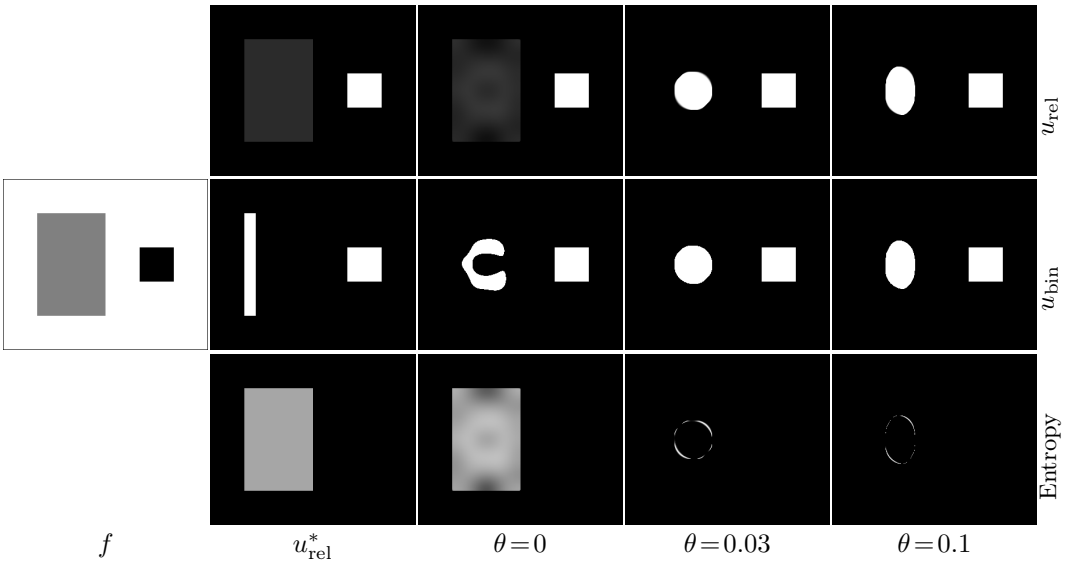


Figure 10.3: Segmenting a fraction of an homogeneous rectangle (gray) with the fixed volume constraint. **Left to right columns:** data term f, analytic optimal solution u^*_{rel} (top row) of the relaxed problem without entropy term. This solution is never perfectly reached due to slow convergence and oscillations. $\theta = 0$ depicts a sample solution. The corresponding rounded solution is far from the optimal solution: a disc with the area of the square next to the square. This optimal solution u^*_{bin} is found by adding the entropy term with $\theta = 0.03$. Increasing the impact of the entropy term to $\theta = 0.1$ increases the non-convexity and adds local optima.

The experiment in Fig. 10.3 demonstrates the relaxation problem with the volume constraint. The goal is to segment a fraction of a homogeneously colored rectangle using the volume constraint. The first column of Fig. 10.3 depicts the data term containing a square with strong foreground preference (black) and a rectangle with equal cost for foreground and background (gray). The white area strongly prefers a background label. The area of the gray rectangle is six times larger than the square. The target volume V_t is chosen to be twice the size of the square, i.e. a sixth of the rectangle shall be filled. Due to the relaxation, the optimal solution is not necessarily compact anymore (u^*_{rel} top row). If no preference is given by the data term or by the boundary conditions, an equal distribution of the volume yields the lowest energy. For the corresponding rounded solution (middle row) we set the first V_t voxels to one according to Prop. 3, but since all pixels in the rectangle

have equal value, their ordering is arbitrary and can lead to many different non-optimal solutions. Since the optimal relaxed solution contains the same information as the data term, optimal rounding is as hard as solving the original problem again in this case. Due to very slow convergence and oscillating behavior of the solver, the optimal relaxed solution is not reached because little volume portions are permanently shifted around and the induced pixel orderings are not related to the optimal binary solution ($\theta = 0$). An optimal binary solution u^*_{bin} - a compact disc with the area of the square - can be optained with the proposed entropy augmentation($\theta = 0.03$ middle row). The entropy term favors binary and thus compact solutions, and ensures that the relaxed solution is close to the binary one which ensures the applicability of simple rounding schemes.

Choice of the entropy weight θ. As a general rule for all experiments in the paper, we found that θ should be chosen as small as possible, but large enough to favor binary solutions. This is because larger θ increase the non-convexity of the problem and thus also the potential number of local optima, as illustrated in the last two columns of Fig. 10.3.

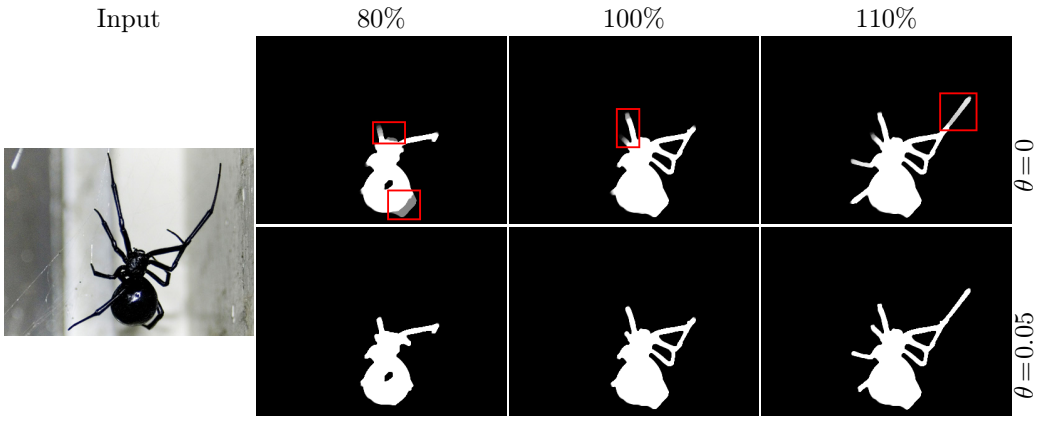


Figure 10.4: Changing the target volume V_t on a real image reveals that relaxation problems occur frequently. The figure compares relaxed solutions u_{rel} with and without entropy term. $V_t \widehat{=} 100\%$ corresponds to the segmentation result without volume constraint.

As shown in Fig. 10.4 the relaxation problems occur frequently and especially if fractions of regions with approximately homogeneous data costs need to be filled. Fig. 10.5 shows two more such cases for which the rounding yields solutions that heavily violate the expected minimal boundary length and how the entropy term avoids these problems.

The right plot in Fig. 10.6 illustrates the rounding scheme and shows non-binary homogeneous regions as plateaus in the sorted label graph which are effectively eliminated by the proposed entropy augmentation. If the target volume seeks into a homogeneous region the rounding might result in shapes not having a minimal contour length (as shown in Fig. 10.5). The left plot of Fig. 10.6 shows that the entropy augmentation consistently yields lower binary energies and leads to smaller energy differences between relaxed and binary solutions which avoids the need for more complex rounding schemes. The relaxed solution of the original problem has always the lowest energy. Conversely, the correspond-

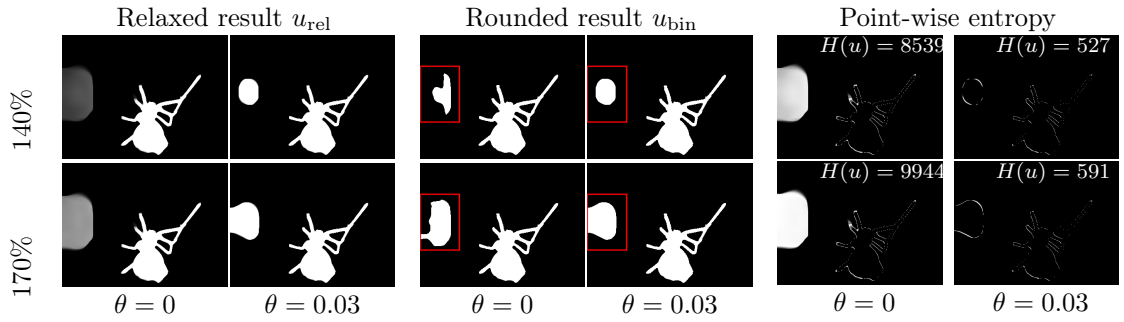


Figure 10.5: Effect of the entropy term on the rounded solution. This figure continues Fig. 10.4 with volume percentages 140% and 170% and shows the strong difference between relaxed and corresponding rounded solutions in regions with approximately homogeneous data costs. The level-sets of u_{rel} in these homogeneous regions do not necessarily obey a minimal boundary length for the enclosed volume (red boxes). The proposed entropy augmentation tackles the problem and ensures that relaxed and binary solutions are more similar.

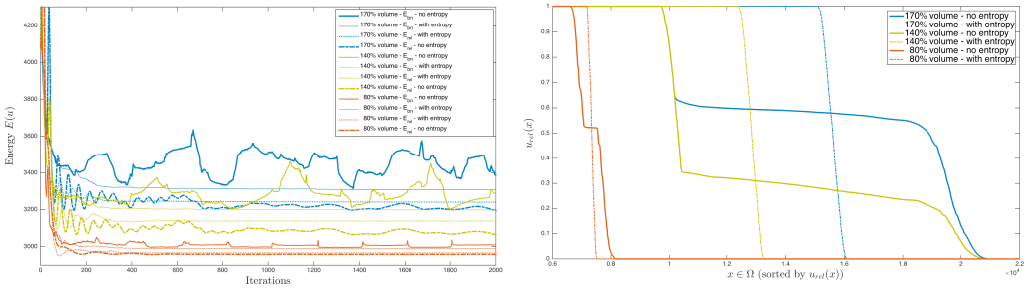


Figure 10.6: This figure further studies the target volumes 80%, 140% and 170% from Figs. 10.4, 10.5. These plots should be read color-wise. Each color represents a different target volume. **Left:** Energy plots during numerical optimization. Binary energies (solid lines) are always larger than relaxed ones (dashed lines). Apart from the first iterations, the energies without entropy term (thick lines) are almost always sandwiching the ones with entropy augmentation (thin lines). That is, the entropy term reduces the gap between relaxed and binary energies. **Right:** Visualization of the rounding scheme (Prop. 3). All pixels are ordered with respect to their relaxed label $u_{\text{rel}}(x)$. The plot shows non-binary pixels 6K to 22K (of 120K). The target volume represents a single point on the x-axis defining the transition between 0 and 1. The non-binary, almost homogeneous image regions form plateaus in this plot and lead to ambiguous selections in the rounding process.

ing rounded energy was always the largest in all our experiments. In all experiments the entropy augmentation stabilized the oscillating behavior of the numerical solver in the presence of non-binary homogeneous image regions and thus lead to better and faster convergence.

10.4.3 Spatio-temporal Multi-View Reconstruction with a Fixed Volume Constraint

The binary 2D image segmentation model from the previous section can be lifted to higher dimensions for spatio-temporal multi-view reconstruction [142]. Then, the fixed volume

constraint can be applied separately to each time frame to express the prior that the overall scene volume should not change over time which is, for instance, approximately true when capturing humans with tight clothing. In this section, we demonstrate 1) that the findings of the previous section are practically even more relevant in a 3D reconstruction setting, and 2) that the proposed entropy augmentation gives consistently better results over entire sequences.

The surface Σ of the reconstructed model is represented as a binary labeling $u : V \times T \to \{0, 1\}$ of interior or exterior defined by the indicator function $u = \mathbf{1}_\Sigma$. It is observed by N cameras with known projections $\{\pi_i\}_{i=1}^N$ and approximate silhouettes $\{S_i(t)\}_{i=1}^N$. The volume-constrained and entropy-augmented reconstruction problem reads

$$\begin{aligned} &\min_u \int_{V \times T} \left[\varrho |D_x u| + g_t |D_t u| + \lambda f u \right] dx + \theta H(u) \\ &\text{s.t.} \int_V u \, dx - V_t = 0 \qquad \forall t \in T \ . \end{aligned} \tag{10.14}$$

Note that we also change the domain from Ω to $V \times T$ in the entropy term $H(u)$ in Eq. (10.5). The regularization term in Eq. (10.14) is split into a spatial and a temporal part. The temporal term is weighted by function $g_t(x, t) = \exp\big(-|\nabla f(x, t)|\big)$ which reduces the temporal smoothing in the presence of motion. The spatial term contains the photoconsistency measure $\varrho(x) : V \times T \to \mathbb{R}_{\geq 0}$ which locally attracts the surface to locations with high photometric consistency, which, in turn, is estimated by means of truncated normalized cross-correlation matching scores of image patches from neighboring cameras. Similarly to the 2D segmentation case, the data fidelity function $f : V \times T \to \mathbb{R}$ gives local preferences for the label of u and is defined as the log-likelihood ratio of the probabilities of being either in the surface interior or exterior. For brevity and readability, we refer to [142] for the exact definitions of the data term and the regularizer weight, also because their influence on the solution is similar to the 2D case. For temporal consistency, three consecutive time frames are jointly solved and longer sequences are processed with a temporal sliding window approach. An iso-surface is extracted from the center frame using the rounding scheme in Prop. 3 and the Marching Cubes algorithm [118].

Fig. 10.7 shows a slice of a single reconstruction together with one of 16 input images, the data term f next to several solutions for different target volumes. The figure compares the impact of the entropy term and shows significant artifacts in the solutions without entropy augmentation. Besides the fact that homogenous cost regions occur frequently in the 3D setting, we observed that noisy cost regions cause the same problems as long they do not infer monotonicity on a larger scale. Especially the Neumann boundary conditions attract the distribution of volume in the entire scene, because it minimizes the regularizer.

Figs. 10.8 and 10.9 show the evaluation of our method on the INRIA dataset [88]. Fig. 10.8 shows energy and volume plots over time. In many frames the volume constraint compensates for low photometric matching scores and distributes the volume according to their score. The energy plot demonstrates the robustness of the entropy augmentation as we consistently obtained lower binary energies for the entire sequence. In Fig. 10.9 we illustrate the benefit of the volume constraint in conjunction with the entropy term. While

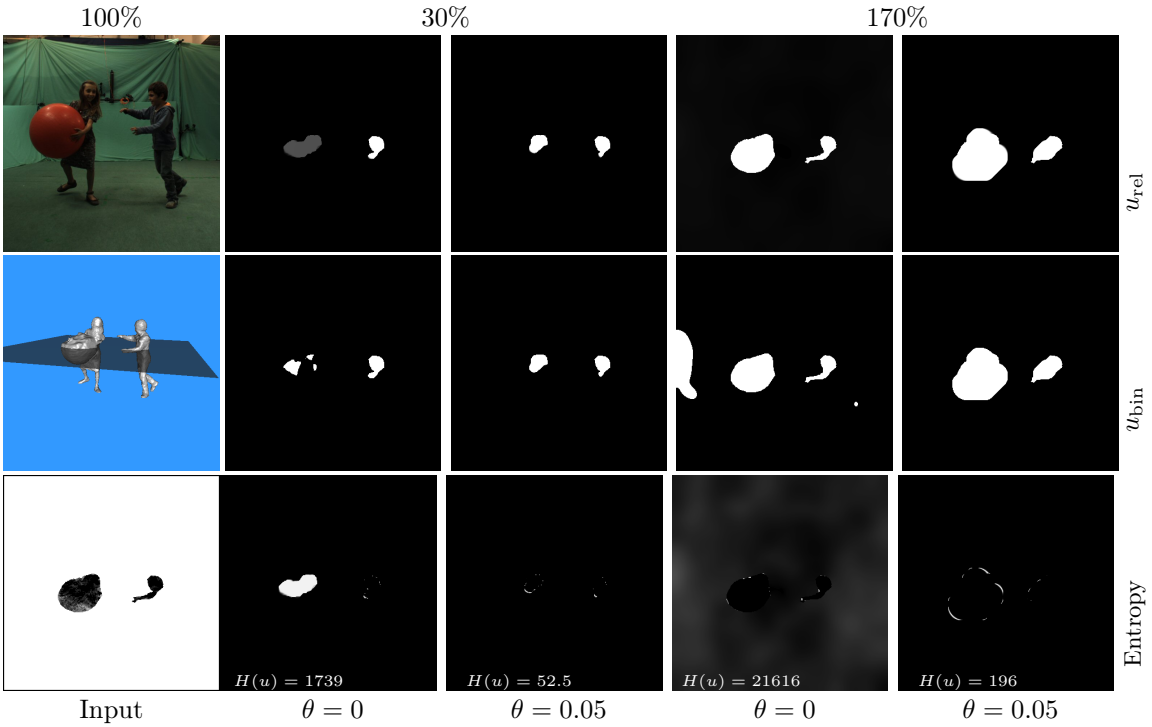


Figure 10.7: Changing the volume of a 3D reconstruction cross section. Homogeneous regions in the cost function occur often in a 3D reconstruction setup which in turn causes non-tight relaxations. The volume adaption can generate strong artifacts which are effectively suppressed with the proposed entropy augmentation.

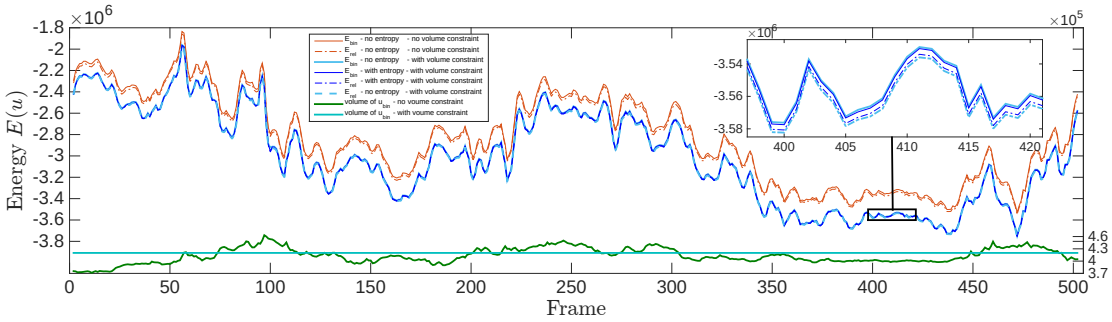


Figure 10.8: Plots of the volume (lower graphs) and several energies (upper graphs) in comparison for 500 frames of a multi-view video sequence (children playing sequence from [88]). Without volume constraint the volume changes over time due to insufficient matching information and occlusions which appear mostly when the volume is above our chosen target volume (cyan line). The upper graphs depict relaxed (dashed) and binary (solid) energies with (blue) and without (red) the volume constraint. Interestingly, the solution without volume constraint always has a higher energy. The blue graphs show binary and relaxed solutions with (light blue) and without (dark blue) the entropy term. For the entire sequence we verified that the solutions without entropy enclose the ones with entropy term, i.e. $E_{\text{rel}}^{\theta=0} < E_{\text{rel}}^{\theta>0} < E_{\text{bin}}^{\theta>0} < E_{\text{bin}}^{\theta=0}$ as shown in the magnified area. That is, our approach gives consistently better binary solutions.

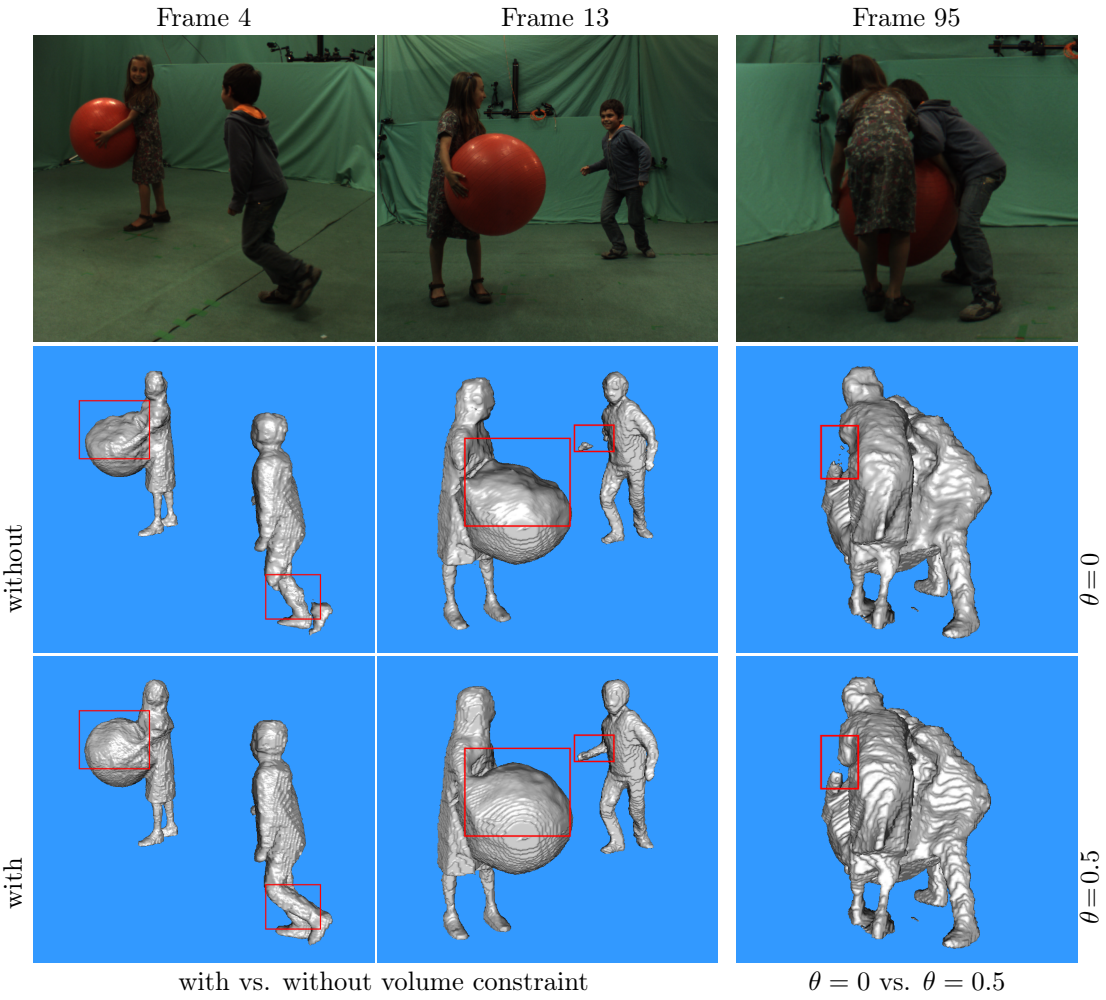


Figure 10.9: Reconstruction results with and without volume constraint (for $\theta = 0.5$) as well as with and without entropy term in comparison (for fixed V_t). For the first 50 frames the volume is too low (see Fig. 10.8 bottom) and the constraint improves the reconstruction (first two columns). Column 3 compares the impact of the entropy term. With entropy term, the regularizer concentrates the volume at locations with a higher data term rather than distributing volume around regions with low data term for attaining smoothness. Therefore, the girls arm is better recovered for the same target volume.

the energy without entropy distributes the volume to generate smooth transitions between opposing labels the entropy term concentrates the volume to the locations with the best likelihood score. In sum, the fixed-volume relaxation problem (10.13) remains tight as long as the data term or the boundary conditions enforce a monotonicity of local costs to avoid fractional homogeneous labelings. The supplementary material provides further details on the experiments.

10.5 Conclusion

We proposed a relaxation technique for general multi-label problems which assures that the computed solutions of the relaxed problem are more binary and consequently have lower optimality gaps. The key idea is to combine the traditional convex relaxations with a concave entropy-term which favors binary solutions. We showed that the arising non-convex problem can be optimized with a provably convergent DC programming method. We demonstrated both theoretically and experimentally that binary solutions are energetically favored and that optimality gaps are smaller. Experiments on multi-region inpainting, image segmentation and spatio-temporal multi-view reconstruction demonstrate that the proposed entropy-based relaxation method is faster and consistently yields solutions of better visual quality and lower energy with respect to the original binary optimization problems.

Chapter 11

Motion Cooperation: Smooth Piece-Wise Rigid Scene Flow from RGB-D Images

---

Authors:	Mohamed Souiai[1]	mohamed.souiai@tum.de
	Mariano Jaimez[1,3]	mariano.jaimez@in.tum.de
	Jörg Stückler[2]	stueckler@vision.rwth-aachen.de
	Javier Gonzalez-Jimenez[3]	javiergonzalez@uma.es
	Daniel Cremers[1]	cremers@tum.de

[1]Technische Universität München, Germany
[2]RWTH Aachen, Germany
[3]University of Màlaga, Spain

Status:	Published
Publication:	Mariano Jaimez , Mohamed Souiai, Jörg Stückler, Javier Gonzalez-Jimenez, and Daniel Cremers. "Motion Cooperation: Smooth Piece-Wise Rigid Scene Flow from RGB-D Images". In: *International Conference on 3D Vision (3DV)*. IEEE, 2015. DOI: 10.1109/3DV.2015.15

Individual contribution	Leading role in realizing the scientific project
Problem definition	significantly contributed
Literature survey	significantly contributed
Implementation	contributed
Experimental evaluation	contributed
Preparation of the manuscript	significantly contributed

Abstract We propose a novel joint registration and segmentation approach to estimate scene flow from RGB-D images. Instead of assuming the scene to be composed of a number of independent rigidly-moving parts, we use non-binary labels to capture non-rigid deformations at transitions between the rigid parts of the scene. Thus, the velocity of any point can be computed as a linear combination (interpolation) of the estimated rigid motions, which provides better results than traditional sharp piecewise segmentations. Within a variational framework, the smooth segments of the scene and their corresponding rigid velocities are alternately refined until convergence. A K-means-based segmentation is employed as an initialization, and the number of regions is subsequently adapted during the optimization process to capture any arbitrary number of independently moving objects. We evaluate our approach with both synthetic and real RGB-D images that contain varied and large motions. The experiments show that our method estimates the scene flow more accurately than the most recent works in the field, and at the same time provides a meaningful segmentation of the scene based on 3D motion.

11.1 Introduction

Scene flow estimation has many applications such as human body pose tracking, articulated object modelling for virtual/augmented reality or traffic scene understanding. In many scenarios, the dynamic scene is composed of rigid parts: human/animal bodies, man-made articulated objects, cars in a street scene, etc. Many existing methods that work on scene flow do not completely exploit this aspect, and estimate motion fields that are only locally rigid or not rigid at all. Other methods do segment the scene to impose rigidity or strong regularization over the regions (or segments). However, these segmentations are only used as tools to improve the accuracy of the estimates, and do not really correspond to the underlying/independent motions of the scene (e.g. [176] partitions the scene into depth layers, [193] divides the scene into piecewise planar regions). Therefore, the segmentation-from-motion problem, which can be particularly useful for scene understanding or human-machine interaction, is not trully addressed by these methods.

On the other hand, assuming purely rigid motions is a strong restriction that is barely fulfilled in organic shapes. When a person moves, there are parts of their body moving rigidly (e.g. upper and lower arms or legs) and others which are transitions between the rigid ones (e.g. the neck). Besides, rigid motions within a fine-grained articulated structure may not be observable with the limited resolution of a camera. For these reasons, a sharp segmentation will never be able to estimate the motion of life beings or some other inanimate objects with exactitude.

In our method, we leverage the natural rigid-part decomposition by allowing for smooth continuous transitions between the parts. We formulate the problem of retrieving a smooth segmentation along with the motion estimates of the rigid parts, where each rigid part is assigned an independent 6 degree-of-freedom motion. To this end, we solve a non-convex optimization problem by means of coordinate descent consisting of a motion estimation step (in the fashion of visual odometry) and a subsequent variational multilabeling solver. By using a weighted quadratic regularizer over the discontinuity-preserving total variation

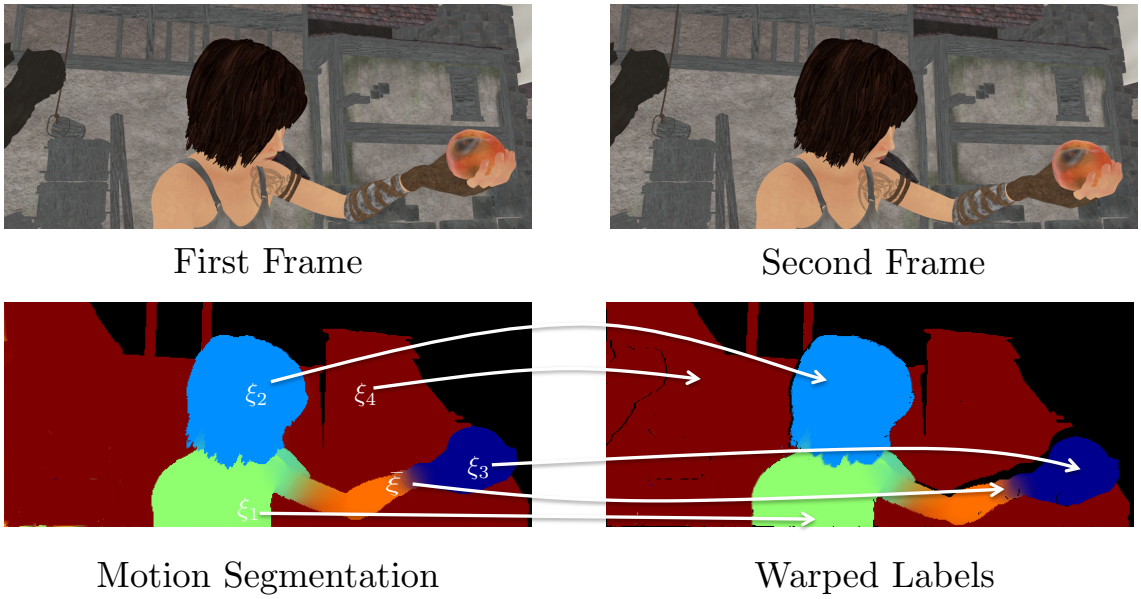


Figure 11.1: The proposed method is based on a motion interpolation model, which allows the emergence of smooth transitions between the segments where the motion is given by a convex combination of adjacent rigid motions (e.g. in $\bar{\xi}$).

(TV), we promote smooth transitions between motion models rather than a harsh competition. For this reason, we refer to this approach as "motion cooperation" as opposed to the traditional "motion competition". We evaluate our motion cooperation scene flow (MC-Flow) algorithm with synthetic and real RGB-D image pairs, and compare it with state-of-the-art approaches. In all cases, our approach achieves a superior performance both qualitatively and quantitatively. Furthermore, this evaluation demonstrates that the combination of a convex relaxation labeling with quadratic regularizer is superior to a sharp traditional segmentation because it naturally relaxes the overly constraining assumption of piecewise rigidity. Additionally, we show that our method retrieves meaningful soft segmentations into rigid parts as depicted in Figure 11.1.

11.1.1 Related work

Scene flow estimation has been traditionally investigated in the multi-view stereo setting within the computer vision community. Vedula *et al.* [190] have proposed one of the first methods based on the optical and range flow constraints. This approach has been later extended to regularize the flow field using quadratic [208] and TV regularization [19, 87], the latter optiziming for disparity and flow jointly. In [197], disparity and scene flow estimation has been decoupled to achieve real-time performance with a stereo camera system. The approach in [192, 193] oversegments the image into superpixels, assumes the superpixels to cover planar regions, and estimates a rigid-body motion for each superpixel individually. In [193], the planar motion of a superpixel acts as a regularization constraint on the scene flow of the individual pixels. Recently, with RGB-D cameras, the scene flow

estimation topic has received further attention due to the availability of depth images at high framerate. Herbst *et al.* [83] used the L1 norm on a data term derived from the optical and range flow constraint equations and showed good qualitative results. Jaimez *et al.* [4] devised the first real-time dense scene flow for RGB-D images. A more natural TV regularization for the flow was proposed, where the regularization term minimizes the line integral of the scene flow gradients over the observed 3D surface. Quiroga *et al.* [153] overparametrize scene flow and estimate a 6-DoF rigid-body motion at each pixel. They regularize the flow field in this 6-DoF parametrization such that their model favors locally rigid motions. Hornacek *et al.* [86] also parametrize the flow-field with 6 DoF, but propose to match corresponding points within a spherical search range instead of traditional planar patch comparisons.

On the other hand, motion segmentation has also been studied in computer vision research. An early variational method for motion segmentation using optical flow constraints was proposed by Cremers and Soatto [48] in their work on motion competition. The name stems from the interpretation of the motion segments to compete for the boundaries through the best fit to their individual motion model. Several extensions to this method have been proposed, e.g. using non-parametric motions [32]. Unger *et al.* [189] explicitly model occlusions as an additional label in the multilabel optimization and impose a map uniqueness constraint to avoid ambiguous (non-bijective) data associations. All these methods are 2D and, hence, do not incorporate a 6-DoF motion model. Furthermore, they estimate a discrete segmentation.

3D-motion segmentation has only gained attention recently, mainly due to the current availability of GPUs and dense RGB-D cameras. Roussos *et al.* [160] propose a variational rigid-body motion segmentation and reconstruction method for monocular video. Zhang *et al.* [207] also pose 3D multi-body structure-from-motion in a variational framework. They require, however, a plane fitting step to make the method robust. Closely related to our method is the approach by Stueckler and Behnke [175]. They jointly estimate motion and segmentation of rigid bodies in an expectation-maximization framework in RGB-D video. Each motion segment is assigned one rigid-body motion, but the approach does not interpolate between the motions of the segments. Recently, Sun *et al.* [176] proposed a probabilistic approach which makes use of a depth-based segmentation to estimate motion between RGB-D images. They regularize the estimation process by retrieving a mean rigid-body motion in each layer. This approach also does not explicitly model smooth transitions of motions between layers, but allows for small deviations of the motion field from the layer's mean motion.

11.1.2 Contributions

The MC-Flow algorithm is the first approach to perform joint soft-labeling and scene flow estimation by dissecting the scene into differently moving regions and their underlying motion. Our contributions are the following:

- Our algorithm estimates 3D motion based on a smooth piecewise rigidity assumption and simultaneously finds a soft motion-based segmentation of the scene.

- By choosing a suitable regularizer we are able to interpolate between rigid motions in order to recover non-rigidly moving parts and their underlying motion.
- An arbitrary (and previously unknown) number of rigid parts can be segmented automatically.
- MC-Flow outperforms state-of-the-art RGB-D scene flow algorithms qualitatively and quantitatively.

11.2 Problem formulation

In this work, we assume that the scene can be segmented into n unknown distinct motion labels, each label standing for one rigid motion, as well as non-rigid parts which can be explained by neighbouring rigid motion labels. An illustration of such a smooth segmentation can be seen in Figure 11.1. As inputs, a pair of RGB-D frames (I_1, Z_1) and (I_2, Z_2) is given, where $I_{(.)} : \Omega \to \mathbb{R}$ and $Z_{(.)} : \Omega \to \mathbb{R}$ stand for the intensity and depth images defined on the image domain $\Omega \subset \mathbb{R}^2$. The segments and the rigid motions associated to them are obtained by minimizing a functional which depends on an implicit labeling function $u : \Omega \to [0,1]^n$, the 6-dimensional twist parametrizations $\xi_i \in \mathbb{R}^6$ of the rigid motions and the number n of rigidly moving parts. The label assigment function u encodes the moving scene in the following way:

$$u_i(x) = \begin{cases} 1 & \text{if } x \in \Omega_i, \\ 0 & \text{if } x \notin \Omega_i, \\ (0,1) & x \text{ belongs partially to } \Omega_i \end{cases} \tag{11.1}$$

Here we denote the i-th segment by $\Omega_i \subset \Omega$, which moves with a velocity ξ_i. Note that, in order to allow for fuzzy assignments, the label functions u_i can take on values in the interval $[0,1]$, in contrast to classical label assignment problems and their underlying binary representation.

The general problem of jointly solving for motion segmentation and motion estimation can be stated as the following optimization problem:

$$\begin{aligned} E_m(\xi, u, n) = &\int_\Omega G(\xi, I_1, I_2, Z_1, Z_2, u, n)\, dx + R(u, n) \\ &s.t. \sum_{i=1}^{n} u_i(x) = 1,\ u_i(x) \geq 0 \quad \forall x \in \Omega \end{aligned} \tag{11.2}$$

The function G encodes geometric and photometric consistency between the RGB-D images according to a linear combination of rigid-body motions:

$$\begin{aligned} G(\xi, I_1, I_2, Z_1, Z_2, u, n) = &F(I_1(x) - I_2(\mathcal{W}_{\bar{\xi}}(x))) \\ &+F(\left|g_{\bar{\xi}}\, \pi^{-1}(x, Z_1(x))\right|_z - Z_2(\mathcal{W}_{\bar{\xi}}(x))) \end{aligned} \tag{11.3}$$

with

$$\bar{\xi} = \sum_{i=1}^{n} u_i(x)\xi_i \quad , \quad \mathcal{W}_\xi(x) = \pi(g_\xi \pi^{-1}(x, Z_1(x)))$$

and $|\bullet|_z$ meaning the z-coordinate. The warping function $\mathcal{W}_\xi(x)$ involves a projection π which transforms the 3D coordinates of the observed points into pixel coordinates. The function g relates twist coordinates to rigid transformation matrices in $SE(3)$. The function F in (11.3) measures photometric / geometric consistency and can be chosen according to the application and prior knowledge. In order to obtain a compact labeling, we regularize the labels by imposing a smoothing term $R(u, n)$ in (11.2). Note that problem (11.2) is hard to minimize because the labels are non-linearly involved in the non-convex dataterm G. To the best of our knowledge, except for performing complete search on u, which is unfeasible in our application, there is no direct way of tackling problem (11.2). Consequently, we consider a simpler formulation where the labels are pulled out of the dataterm. This significantly facilitates the optimization process because the label assignment function u is now linearly involved with the dataterm:

$$\begin{aligned} E_r(\xi, u, n) = \sum_{i=1}^{n} \int_\Omega u_i D(\xi_i, I_1, I_2, Z_1, Z_2)\, dx + R(u, n) \\ s.t. \sum_{i=1}^{n} u_i(x) = 1,\ u_i(x) \geq 0 \quad \forall x \in \Omega \end{aligned} \tag{11.4}$$

The data fidelity term D_i is now evaluated for every independent rigid motion:

$$\begin{aligned} D(\xi_i, I_1, I_2, Z_1, Z_2) = F(I_1(x) - I_2(\mathcal{W}_{\xi_i}(x))) \\ + F(\left|g_{\xi_i}\, \pi^{-1}(x, Z_1(x))\right|_z - Z_2(\mathcal{W}_{\xi_i}(x))) \end{aligned} \tag{11.5}$$

The optimization problems (11.2) and (11.4) would be equivalent if the labels u were binary. The main difference between the two models is that in (11.2) the motions are interpolated and subsequently used to evaluate the residuals with the exact velocities, whereas in (11.4) the residuals are computed for each independent rigid motion and interpolated afterwards. With binary labels, there would not be interpolation between motions or residuals and, hence, both models would turn out to be the same. In this work, we aim to solve the motion interpolation model (11.2) but, given its complexity, we resort to the simpler model (11.4) as an approximation of (11.2) to optimize for the labels. For this reason, the regularization term $R(u, n)$ plays a crucial role to estimate accurate interpolated motions at the transitions between rigid bodies/parts.

11.2.1 Overall Optimization

Independently of which of the two models we chose, the dataterms are nonlinear with respect to the rigid motions. Therefore, the overall optimization problem is not convex and the global minimum cannot be guaranteed to be found.

To tackle this joint problem, we propose a coordinate descent strategy that alternates between estimating the motions for a fixed set of labels and then refining these labels for the recently obtained velocities, as illustrated in Algorithm 6. The motions are computed in the fashion of a visual odometry problem, but considering that the whole scene is not rigid but smooth-piecewise rigid. The labels are solved using the approximate model (11.4) that is convex in u. Note that we are implicitly optimizing for the label count n by adapting the number of labels within the inner iterations, as will be described in section 11.5. Next, we elaborate on how to solve the main two subproblems in Algorithm 6.

Algorithm 6 Coordinate Descent Optimization for joint Motion Estimation and Segmentation

Initialize u^0

1: **for** $k = 0, 1, 2, ...$
   $\xi^{k+1} = \arg\min_{\xi} E(\xi, u^k)$
   $u^{k+1} = \arg\min_{u} E(\xi^{k+1}, u)$
   Update n
2: **end for**

11.3 Motion estimation

Given a precomputed set of labels, at every iteration of Algorithm 6 we need to estimate the rigid-body motions associated to each label (step 1). This problem can be considered as an extension of the well-known visual odometry (VO) problem. In this more general case, the whole scene is not supposed to be moving rigidly; instead, we assume that there are n predominant rigid motions that can be linearly combined to explain the motion of every point of the scene.

Our solution to estimate the motion of the segments builds upon two existing VO methods: DIFODO [89] and the Robust Dense Visual Odometry [94]. This solution is obtained by minimizing the photometric and geometric residuals, defined as

$$r_I(x) = I_1(x) - I_2(\mathcal{W}_{\bar{\xi}}(x)) \tag{11.6}$$

$$r_Z(x) = \left| g_{\bar{\xi}}\, \pi^{-1}(x, Z_1(x)) \right|_z - Z_2(\mathcal{W}_{\bar{\xi}}(x)) \tag{11.7}$$

Note that the residuals are defined here according to the motion interpolation model (11.2). To cope with large motions, the process of minimization is applied in a coarse-to-fine scheme where the residuals are linearized at each level of the pyramid. In order to deal with outliers and to provide an accurate motion estimate, a robust function of the residuals is minimized:

$$\xi = \arg\min_{\xi} \left\{ \int_{\Omega} F(r_I) + \alpha F(r_Z) dx \right\} \tag{11.8}$$

$$F(r) = \frac{c^2}{2} ln \left(1 + \left(\frac{r}{c} \right)^2 \right) \tag{11.9}$$

The function F is equivalent to the Cauchy M-estimator. Although we do not present comparisons in this regard, it was chosen because it provides considerably better results than other more common choices like the L2 or L1 norms. The parameter α balances the two kinds of residuals and c controls the relative weighting between high and low residuals. This minimization problem is solved using Iteratively Reweighted Least Squares (IRLS), where the associated weighting function is

$$w(r) = \frac{1}{1 + \left(\frac{r}{c}\right)^2}. \tag{11.10}$$

With this strategy, we are able to solve the motion estimation problem accurately. The minimization of both the photometric and the geometric residuals allows us to estimate the motion of the segments even if they lack of texture or geometric distinctive features. This aspect is crucial because the segments can be considerably small (compared to the whole scene) and might not present sufficient photometric or geometric data to solve the 3D registration problem using only one of these two input data.

11.4 Label optimization

Once the motion ξ^{k+1} at a given iteration $k+1$ is obtained, we optimize the label assignment function as the second step of the overall optimization problem (Algorithm 6). For a fixed set of motions ξ, the functional $E(\xi^{k+1}, u)$ is convex and can be solved using state-of-the-art first-order solvers. In this work, the labeling function is optimized with the primal-dual algorithm developed by Pock *et al.* [150]. Detailed information about how to apply this algorithm to the addressed problem is given in the supplementary material.

In this work, two different regularizers are considered: total variation and quadratic regularization. Furthermore, the geometrical data that RGB-D cameras provide are exploited to regularize the labels according to the real 3D distances between points. Thus, regularizers are defined as a function of a weighted gradient ∇_r of the labels, whose weights (r_x) are the inverse of the 3D distances between the points:

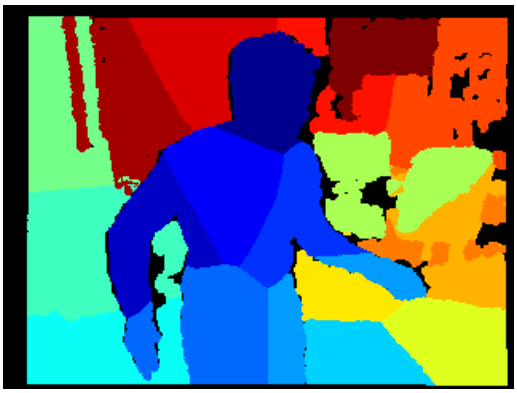
$$\nabla_r u_i = \left(r_{x_1} \frac{\partial u_i}{\partial x_1}, r_{x_2} \frac{\partial u_i}{\partial x_2} \right) \tag{11.11}$$

More details on the theory and the implementation of this regularization strategy can be found in [4].

11.4.1 Total Variation Regularization

Total variation was made popular by the seminal work of Rudin Osher Fatemi (ROF) [161] on image denoising. The most prominent properties of the TV regularizer are allowing for jumps in the solution and being a measure of perimeter of a region if applied on its indicator function. These factors made TV widely used in general reconstruction problems

a) Input image



b) K-means initialazation u^0

Figure 11.2: We initialize our algorithm by performing K-means (k=20) on the 3D coordinates of the image pixels.

like image denoising [161], image deblurring [42] and image segmentation [38, 138]. In order to incorporate TV regularization into our approach, we simply set:

$$R(u,n) = \lambda \sum_{i=1}^{n} \int_{\Omega} \|\nabla_r u_i(x)\|_1 \, dx \tag{11.12}$$

11.4.2 Quadratic Regularization

As previously mentioned, TV regularization favors sharp label boundaries. However, in our segmentation we would like to obtain a smooth interface between the labels. Hence, a suitable choice to encourage smooth label transitions is the so-called Tikhonov or quadratic regularization:

$$R(u,n) = \lambda \sum_{i=1}^{n} \int_{\Omega} \|\nabla_r u_i(x)\|_2^2 \, dx \tag{11.13}$$

Normally, quadratic regularization does not allow for discontinuities in the solution, which would not help to provide a precise segmentation. However, the geometric weighting makes it able to estimate discontinuities in the labels and soft transitions between rigid parts at the same time.

11.5 Initialization and adaptive number of labels

This section describes the adopted strategy to refine the number of labels n so that they represent the actual number of independent rigid motions in the scene. Since we are solving a non-convex problem, it is crucial to start with an initial set of labels u^0 that allows us to converge to the global optimum in Algorithm 6. Instead of including the number of labels in the variational formulation (which would significantly increase the computational burden), we propose to initialize the labels with a meaningful over-segmentation of the

observed scene and iteratively remove those labels that are redundant or not significant for the overall motion estimation. To this end, we create an initial K-means segmentation based on the 3D coordinates of the points of the scene. The initial number of labels is always set to 20 (the number of independent rigid motions in the scene is assumed to be smaller than this quantity). An example of a K-means initialization is shown in Figure 11.2. The refinement of the label count is performed after a full inner iteration of Algorithm 6 as follows:

- If labels i and j are associated to similar velocities, i.e., if $\|\xi_i - \xi_j\| \leq \delta$ for some small $\delta > 0$, we merge both labels.
- If a label i contains too few pixels, i.e., if $\int_\Omega u_i(x) < \gamma$ for some small $\gamma > 0$, we assign these pixels to the outlier label and remove label i.

11.6 Occlusions and outliers

In our formulation, we include an outlier label (u_n) to capture pixels with null depth measurements and those other pixels that produce very high residuals for all the possible velocity candidates ξ_i. To this end, a constant weight K_D is associated to this label which, according to (11.4) means that $D_n = K_D$ in the whole image plane Ω. As previously mentioned, this outlier label also plays an important role in the process of reducing the number of labels. When a label is removed as a consequence of containing very few pixels, those few pixels need to be assigned to another label. If they were assigned to a wrong label they could affect the subsequent motion estimate and spoil the results. Conversely, if they are assigned to the outlier label, they don't participate in the motion estimation stage and are automatically assigned to the best label afterwards in the label optimization stage.

On the other hand, we detect occlusions to avoid the evaluation of the dataterm (D_i in (11.4)) for those pixels which are not visible in the second RGB-D frame. Occlusions are handled with a binary mask $O(x)$ instead of an extra label, in a way that occluded points can still be segmented and, therefore, their 3D motion is estimated too. This can be accomplished by virtue of the regularization term, and allows us to provide a complete segmentation of the scene even if some points or areas are occluded after the motion.

In order to detect occlusions, two factors are considered: the amount of pixels that are registered to each pixel of the second frame and the temporal change in the depth images. First, we compute a cumulative function $C(x) : \Omega \in \mathbb{R}^2 \rightarrow \mathbb{R}$ that counts how many pixels from the first frame are warped to the pixel x of the second frame (according to the estimated motion). Without occlusions, this function is approximately equal to 1 (or maybe inferior to one for new points appearing in the second frame), meaning that there is a one-to-one (bijective) correspondence between the observed points at both images. On the contrary, if $C(x)$ is noticeably higher than one, there are some pixels in the first frame that are warped to the same pixel x in the second frame, indicating the existence of occlusions. Consequently, we can define a function $O_C(x)$ that finds the pixels candidates for occlusion by applying a warping with the estimated motion and evaluating the cumulative function

C:

$$O_C(x) = C(\mathcal{W}_{\bar{\xi}}(x)) \tag{11.14}$$

On the other hand, unlike in the optical flow problem, geometric information is available and can be exploited to reason whether a point is occluded or not. The simplest function that can be used to detect occlusions is the temporal change in depth:

$$O_Z(x) = Z_1(x) - Z_2(x) \tag{11.15}$$

Combining these two functions we can detect most of the occluded areas in the scene by imposing a threshold K_o:

$$O(x) = \begin{cases} 1 & \text{if} \quad O_C(x) + K_z O_Z(x) > K_o \\ 0 & \text{else} \end{cases} \tag{11.16}$$

where K_z is a parameter that weights O_Z against O_C. This strategy could be improved by embedding these functions into a variational formulation and imposing regularization over the occlusion mask. However, this has not been implemented in our work because it would significantly increase the runtime of our method.

11.7 Experiments

In this section, qualitative and quantitative results are presented to evaluate the accuracy of our approach. These results are divided into two categories: scene segmentation and scene flow estimation. However, the evaluation process is not straightforward given the lack of benchmarks with either scene flow ground truth or segmentation from 3D motion. For this reason, we have selected a set of synthetic and real RGB-D frame pairs that

	Photometric residual - RMSE					Geometric residual - RMSE				
	PD-Flow [1]	SR-Flow [3]	Layered-Flow [4]	MC-TV	MC-Quad	PD-Flow	SR-Flow	Layered-Flow	MC-TV	MC-Quad
Sintel-1	0.060	0.035	0.049	0.022	**0.021**	0.443	0.317	0.420	0.253	**0.186**
Sintel-2	0.057	0.068	0.063	0.026	**0.025**	0.086	0.090	0.108	0.056	**0.053**
Sintel-3	0.048	0.041	0.047	0.032	**0.028**	0.021	0.022	0.035	0.018	**0.017**
Sintel-4	0.091	0.069	0.109	0.063	**0.044**	0.378	0.347	0.607	**0.155**	0.190
Sintel-5	0.074	0.067	0.091	**0.051**	0.055	0.373	0.267	0.498	**0.203**	0.283
Sintel-6	0.120	0.118	0.127	**0.055**	**0.055**	0.224	0.190	0.253	0.114	**0.096**
Sintel-7	0.076	0.071	0.079	**0.035**	0.038	0.407	0.423	0.382	0.233	**0.188**
Sintel-8	0.063	**0.026**	0.045	0.028	0.027	0.083	0.069	0.086	0.038	**0.037**
RI-1	0.038	0.025	0.031	0.024	**0.022**	0.070	0.060	0.046	**0.038**	**0.038**
RI-2	0.032	0.028	0.035	0.021	**0.020**	0.286	0.259	0.294	0.114	**0.102**
RI-3	0.031	0.024	0.027	**0.018**	**0.018**	0.221	0.208	0.217	0.160	**0.145**
RI-4	0.015	0.012	0.011	**0.008**	**0.008**	0.025	**0.024**	0.025	0.025	0.025
RI-5	0.074	0.051	0.056	**0.039**	0.040	0.095	0.087	0.108	**0.079**	0.085
RI-6	0.077	0.050	0.070	0.049	**0.047**	**0.036**	0.038	0.037	0.041	0.040
Average	0.061	0.049	0.060	0.034	**0.032**	0.197	0.172	0.223	0.109	**0.106**

Table 11.1: Photometric and geometric residuals after warping the image pairs with the estimated scene flow.

contain varied and challenging motions. First, our approach is tested with some sequences from the Sintel dataset [34]. This dataset contains scenes with heterogeneous and large motions, and provides optical flow ground truth which can be used to measure the scene

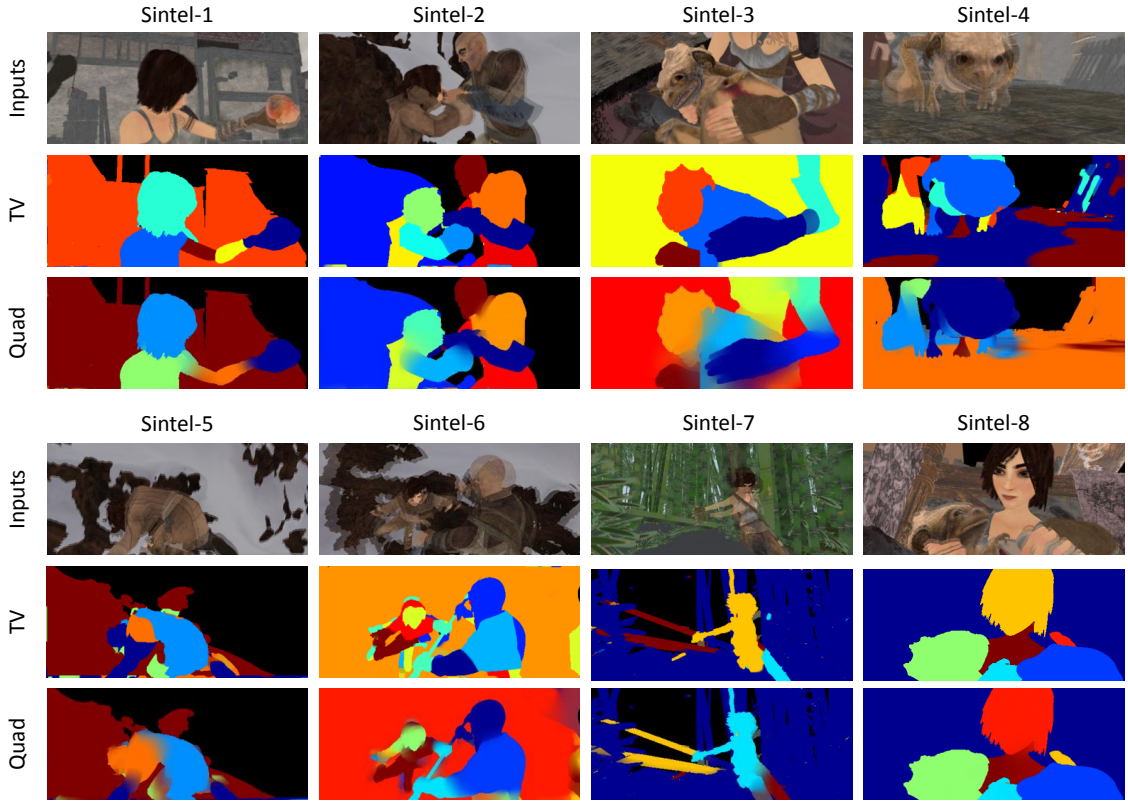


Figure 11.3: Segmentation estimated by our approach for the eight sequences of the Sintel dataset considered. Colors are independent for each result and do not depend on the associated rigid motion. Black represents the outlier label.

flow error. Second, the joint segmentation and motion estimation is generated for several RGB-D image pairs that either have been utilized in previous works in the literature (as in [153]) or have been taken with RGB-D cameras in our lab. In all cases, two versions of our method are tested, corresponding to the two different regularization strategies for the label optimization problem: total variation (TV) and quadratic regularization (Quad). The resolution adopted for the images is QVGA (240×320) for those taken with an RGB-D camera and 218×512 for the Sintel sequences. The maximum depth is set to 5 meters in all cases. Tests have been performed with a total of fourteen image pairs: eight from the Sintel dataset (named "Sintel-1...8") and six real image pairs (named "RI-1...6").

11.7.1 Scene segmentation

In this subsection we present the motion segmentation that our method provides for all the tested sequences. The occlusion layer is also displayed for some sequences together with the segmentation although the occlusion is not a label itself (but a mask). Figure 11.3 shows the results for the Sintel images. It can be observed that TV produces very sharp labels with very few pixels interpolating between different motions. On the contrary,

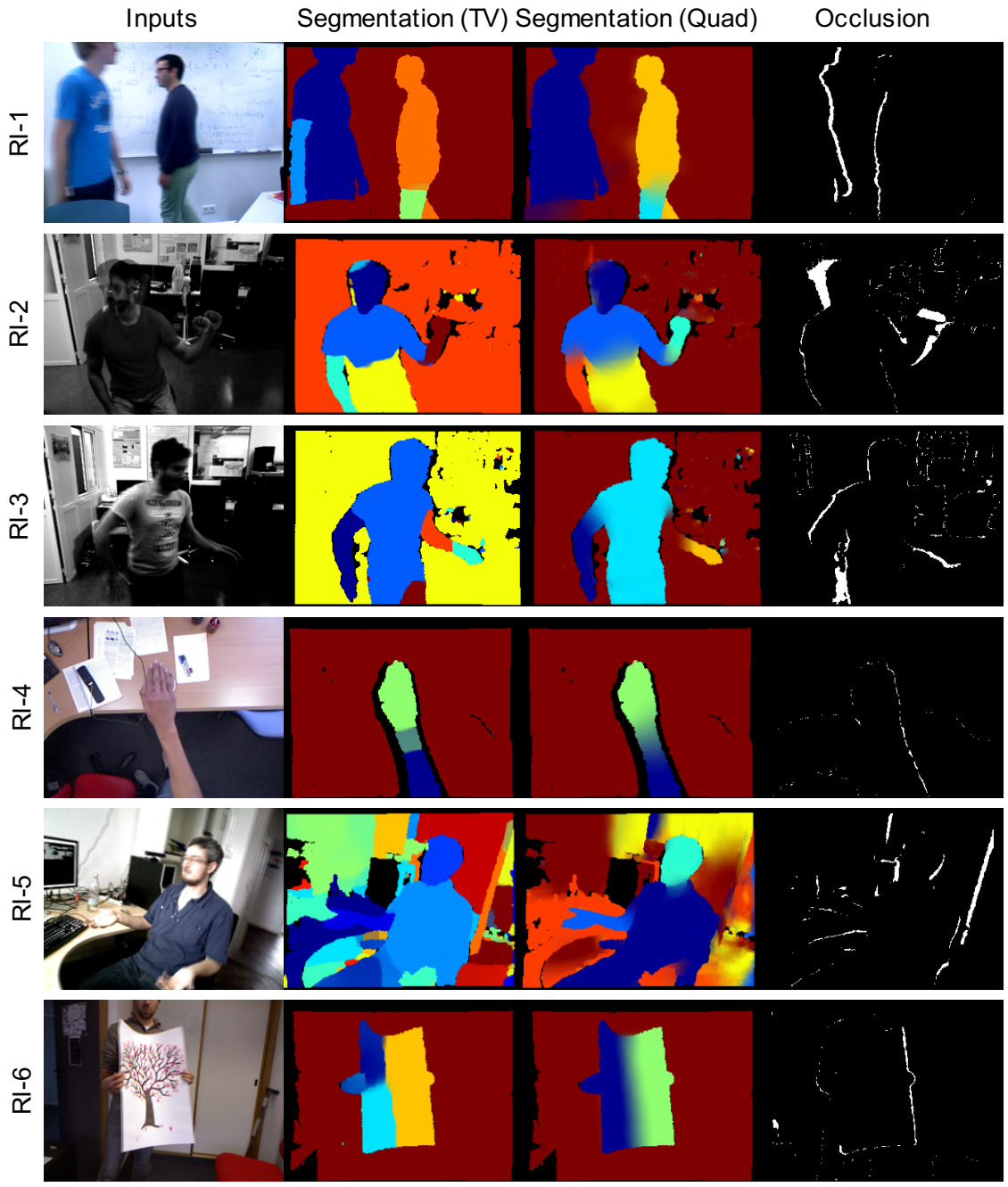


Figure 11.4: Segmentation and the occlusion layer estimated by our approach for 6 image pairs taken with RGB-D cameras. Colors are independent for each result and do not depend on the associated rigid motion. Black represents the outlier label.

quadratic regularization gives rise to a smooth segmentation where many pixels adopt an interpolated velocity between two (or maybe more) rigid-body motions. The same behavior can be seen in Figure 11.4 where the results for the real RGB-D images are presented. In

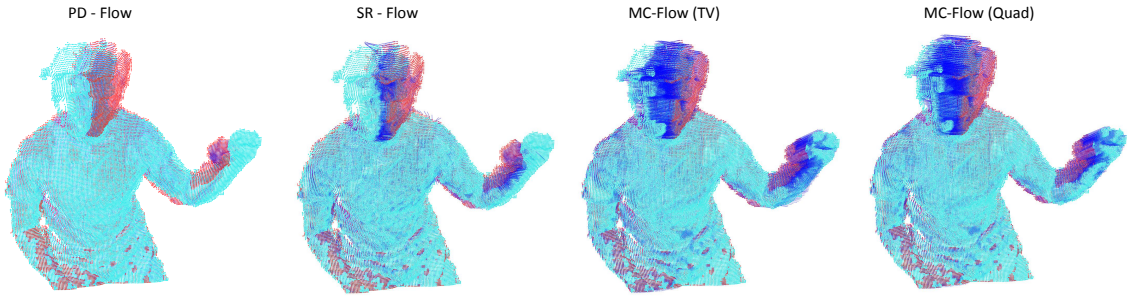


Figure 11.5: Comparison of the 3D motion fields estimated for the "RI-2" sequence. The initial frame is represented by the red point cloud, the final frame by the turquoise point cloud and the scene flow by the blue lines. The above comparison shows that our approach provides the most accurate estimate of the scene flow.

	Optical flow - EPE					Optical flow - AAE				
	PD-Flow [1]	SR-Flow [3]	Layered-Flow [4]	MC-TV	MC-Quad	PD-Flow	SR-Flow	Layered-Flow	MC-TV	MC-Quad
Sintel-1	1.940	0.684	1.320	0.221	**0.219**	27.87	7.694	13.26	**2.486**	2.827
Sintel-2	2.299	2.100	2.851	0.367	**0.324**	23.63	16.02	35.50	**4.826**	4.950
Sintel-3	1.223	1.130	0.975	0.383	**0.344**	31.69	20.21	20.80	8.364	**7.721**
Sintel-4	17.04	21.68	15.26	10.23	**3.436**	73.57	90.56	43.09	22.13	**9.694**
Sintel-5	4.381	3.990	3.212	2.316	**1.983**	24.27	26.14	10.43	14.56	**10.16**
Sintel-6	6.045	7.739	7.67	**1.168**	1.498	12.10	18.99	27.52	**3.845**	5.194
Sintel-7	2.875	3.335	3.382	**1.480**	1.591	26.50	21.26	22.48	**7.723**	8.169
Sintel-8	1.674	0.456	1.012	**0.228**	**0.228**	22.45	4.713	8.003	3.762	**3.757**
Average	4.685	5.142	4.461	2.049	**1.203**	30.26	25.70	22.63	8.462	**6.559**

Table 11.2: Average end-point and angular errors of the optical flow computed by projecting the estimated scene flow onto the image plane.

general, it can be noticed that the number of labels to which the method converges is not the same for the two regularization strategies. Normally, TV produces a higher number of labels because it is not able to interpolate motions and tends to keep extra labels to compensate for it. It can be observed that, but for Sintel-4 (with TV) and RI-5, the resulting segmentations represent quite accurately the different objects and rigid parts of the scenes.

11.7.2 Scene flow evaluation

For all the sequences, the scene flow is evaluated quantitatively and compared with three state-of-the-art methods: the Primal-Dual flow (PD-Flow) [4], the Semi-Rigid flow (SR-Flow) [153] and the Layered flow [176]. First, the photometric and geometric residuals are computed by warping the intensity and depth images (respectively) according to the estimated flow. It is important to note that occluded pixels will show very high residuals even if the motion is accurately estimated for them, which considerably disturbs the error metrics (RMSE of the residuals). To overcome this limitation and to provide more precise comparisons, we compute the RMSE of the non-occluded pixels, which is a more reliable metric of the scene flow accuracy. To this end, we assume that the occlusion layer computed by our approach is sufficiently accurate and use it in all cases (neither PD-flow nor SR-flow detect occlusions). This does not represent any bias toward our method because it

is a common mask applied to all of them, and if some occluded pixels have not been detected properly then they will affect the error metrics of all the compared methods equally. Table 11.1 shows the results for all the frame pairs. It can be observed that our method provides the most accurate estimates with both TV and quadratic regularization. The differences between TV and Quad are essentially caused by the way they produce transitions between the labels and the number of labels they converge to. As previously analyzed, TV generates a sharp segmentation where the motion is barely interpolated, whereas quadratic regularization provides smooth transitions between the labels that lead to larger areas with interpolated motions. On the other hand, TV tends to converge to a higher number of labels, which helps to compensate for its inability to capture nonrigid motions. Overall, the best results are obtained with quadratic regularization, although the differences are small.

For the sake of clarity, Figure 11.5 is included to illustrate the 3D motion field that the compared methods estimate for the sequence "RI-2". PD-Flow, which was conceived to work in real-time, is unable to estimate large motions and can only capture the motion of the body and the upper arms. SR-Flow provides better results but is still unable to reproduce the real motion of the hand and head. Only our approach estimates the whole motion field properly, specially with quadratic regularization of the labels.

Moreover, for the Sintel image pairs, we project the scene flow onto the image plane to obtain the optical flow and compare it with the ground truth provided by the Sintel dataset. In this case we evaluate two error metrics: the average end-point error (EPE) and the average angular error (AAE), as explained in [18]. Again, the results (Table 11.2) are computed for the non-occluded pixels, which is a fairer comparison given that some methods do not manage occlusions and hence provide bad estimates for the occluded areas. It can be seen that our approach with both TV and quadratic regularization clearly outperforms the others, providing a motion estimate that is between 2 and 5 times more accurate than those from the PD-Flow, SR-Flow and the Layered-Flow.

Regarding the computational performance, our method ranks second with a runtime of 30 seconds. For the experiments, we have utilized a standard desktop PC running Ubuntu 14.04 with an AMD Phenom II X6 1035T CPU at 2.6 GHz, equipped with an NVIDIA GTX 780 GPU with 3GB of memory. The measured runtimes are:

- PD-Flow: 0.042 seconds (GPU).
- SR-Flow: 150 seconds (CPU).
- Layered-Flow: 8 minutes (CPU).
- MC-Flow: 30 seconds (label optimization on GPU and all the remaining steps on CPU).

11.8 Conclusion

In this paper we have addressed the problem of joint segmentation and scene flow estimation from RGB-D images. The overall optimization problem is solved by means of a coordinate

descent method which alternates between motion estimation and label optimization, while at the same time adapts the number of labels to the real number of independent rigid motions of the scene. Two different regularization strategies for the labels are employed, TV and quadratic, leading to sharp and smooth segmentations, respectively. Our method has been tested with both synthetic and real RGB-D image pairs, and the experiments show that joint segmentation and motion estimation provides very accurate results that outperform state-of-the-art scene flow algorithms on RGB-D frames. Comparisons between the two regularization strategies show that quadratic regularization estimates motion more accurately than TV because it generates smooth label transitions between rigid bodies, which models the scene motion more realistically. For future work, we plan to extend this work to RGB-D video streams where temporal regularization can be imposed.

Part III

Conclusion and Outlook

Chapter 12

Summary

In this thesis, we made contributions in several areas including:

- Continuous semantic multi-labeling
- Convex relaxation methods
- Motion estimation and segmentation

We consider these disciplines to be crucial towards holistic scene understanding. In large part, this thesis dealt in large part with introducing convex semantic priors for continuous semantic multi-labeling. In addition, we proposed a framework which promotes more binary solutions for convex relaxation methods. Finally, we developed an algorithm for joint motion estimation and image segmentation.

Semantic Priors

This thesis dealt, among other things, with different *convex semantic priors for continuous multi-label segmentation*. These include:

- Co-occurrence priors [10] which we introduced into the framework of continuous multi-label optimization. Using co-occurrence statistics in multi-labeling improves on the segmentation results since unlikely combinations of labels are penalized. This publication is included in Chapter **??**.
- Hierarchical priors [8] which exploit contextual information in order to refine the labeling. Compared to methods using co-occurrence priors, this approach is more principled and the algorithm is no longer dependent on learning label statistics from databases. The problem is cast in a single convex cost function and the optimization is performed jointly with respect to label and scene variables. We included this publication in Chapter 9.
- Generalized minimal description length priors. In Chapter 6, we presented a generalization of the classical minimal description length prior by compositions with an arbitrary monotone convex functions. This enables us to impose, in addition

to merely penalizing the label count, an upper bound on the number of occurring labels. We adopted this constraint and devised a convex approximation of the piecewise constant Mumford-Shah model.

The above approaches are based on energy minimization and have in common that they can be cast in a single convex optimization problem. In Chapter 6, we presented a *unifying framework for semantic image segmentation* which encompasses these priors. Within this framework, we derived a generic primal-dual algorithm for semantic multi-label optimization.

Entropy Minimization for Mixed-Integer Programming

In Chapter 7, we presented a *method for solving mixed-integer programs*. For this purpose, we augment classical convex relaxation strategies with an entropy term which favors integer solutions. We solved the arising convex-concave problem using a provably convergent DC algorithm. We applied our approach on the problem of image cartooning which can be cast in a mixed-integer program. In Chapter 10, we included a publication [9] in which we proposed a similar strategy for improving the integrality of the solution of convex relaxed multi-labeling problems.

Joint Motion Estimation and Segmentation

In Chapter 11, we included our research paper on joint image segmentation and motion estimation [5]. The problems of identifying objects and estimating their motion are highly correlated and very important for a holistic scene understanding. Our approach is able to jointly reason about image segments and their underlying rigid body motion by means of energy minimization. Additionally, by allowing smooth label transitions, we are able to recover non-rigidly moving parts.

Chapter 13

Future Work

In the following, we would like to point out possible directions for further research:

- **Combining semantic segmentation with motion estimation**. In this thesis, we presented a method which segments the scene solely based on its underlying rigid body motions. A possible future work would be jointly performing motion estimation, image segmentation and reasoning about the type of the moving objects. By this, we hope improving the tasks of object recognition, motion estimation and segmentation since this should helps further resolving ambiguities inherent to these problems.
- **Adapting the entropy parameter**. In this thesis, we presented augmenting mixed-integer programs with an entropy term. A future work could be increasing gradually the associated parameter during the optimization. This means that we warm start the algorithm by first solving the convex relaxed problem and increasing the influence of the entropy penalty. Another promising idea is treating this parameter as a Lagrange multiplier associated with an entropy constraint. By solving the Lagrangian problem the adaptation of this parameter comes natural in form of an update step.
- **DC-Programming for non-convex regularizers**. Although established in the optimization community, DC-Programming did not grow famous in the area of computer vision. By decomposing non-convex regularizers into a convex and concave part, Gasso et al. [64] devised a convergent algorithm, based on DC-Programming, for several sparsity promoting penalties. A promising idea is applying this algorithm to variational problems in computer vision in order to implement non-convex regularizers.

Own Publications

[1] J. Bergbauer, C. Nieuwenhuis, M. Souiai, and D. Cremers. «Proximity Priors for Variational Semantic Segmentation and Recognition». In: *IEEE International Conference on Computer Vision (ICCV), Workshop on Graphical Models for Scene Understanding*. 2013.

[2] V. Golkov, M. Menzel, T. Sprenger, M. Souiai, A. Haase, D. Cremers, and J. Sperl. «Direct Reconstruction of the Average Diffusion Propagator with Simultaneous Compressed-Sensing-Accelerated Diffusion Spectrum Imaging and Image Denoising by Means of Total Generalized Variation Regularization». In: *International Society for Magnetic Resonance in Medicine (ISMRM) Annual Meeting*. 2014 (cited on p. 26).

[3] V. Golkov, M. Menzel, T. Sprenger, M. Souiai, A. Haase, D. Cremers, and J. Sperl. «Improved Diffusion Kurtosis Imaging and Direct Propagator Estimation Using 6-D Compressed Sensing». In: *Organization for Human Brain Mapping (OHBM) Annual Meeting*. 2014 (cited on p. 26).

[4] M. Jaimez, M. Souiai, J. Gonzalez-Jimenez, and D. Cremers. «A Primal-Dual Framework for Real-Time Dense RGB-D Scene Flow». In: *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*. 2015 (cited on pp. 26, 94, 98, 104).

[5] M. Jaimez, M. Souiai, J. Stückler, J. Gonzalez-Jimenez, and D. Cremers. «Motion cooperation: smooth piece-wise rigid scene flow from rgb-d images». In: *International Conference on 3D Vision (3DV)*. IEEE. 2015, pp. 64–72 (cited on pp. 10, 24, 26, 61, 110).

[6] Y. Kee, M. Souiai, D. Cremers, and J. Kim. «Sequential Convex Relaxation for Mutual-Information-Based Unsupervised Figure-Ground Segmentation». In: *Proc. Int. Conference on Computer Vision and Pattern Recognition (CVPR)*. 2014 (cited on p. 26).

[7] C. Kerl, M. Souiai, J. Sturm, and D. Cremers. «Towards Illumination-invariant 3D Reconstruction using ToF RGB-D Cameras». In: *International Conference on 3D Vision (3DV)*. 2014.

[8] M. Souiai, C. Nieuwenhuis, E. Strekalovskiy, and D. Cremers. «Convex optimization for scene understanding». In: *IEEE International Conference on Computer Vision (ICCV), Workshop on Graphical Models for Scene Understanding (GMSU)*. 2013 (cited on pp. 8, 22, 24, 25, 39, 43, 59, 109).

[9] M. Souiai, M. R. Oswald, Y. Kee, J. Kim, M. Pollefeys, and D. Cremers. «Entropy minimization for convex relaxation approaches». In: *Proc. Int. Conference on Computer Vision (ICCV)*. accepted. Santiago, Chile, 2015 (cited on pp. 8, 19, 24, 26, 48, 49, 55, 60, 110).

[10] M. Souiai, E. Strekalovskiy, C. Nieuwenhuis, and D. Cremers. «A Co-occurrence Prior for Continuous Multi-Label Optimization». In: *Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR)*. 2013 (cited on pp. 8, 21, 22, 24, 25, 39, 43, 59, 62, 66, 109).

[11] M. Souiai, E. Strekalovskiy, C. Nieuwenhuis, and D. Cremers. «Label Configuration Priors for Continuous Multi-Label Optimization». In: *Technical report*. 2013 (cited on pp. 21, 22, 39, 43, 44).

[12] F. Stangl, M. Souiai, and D. Cremers. «Performance evaluation of narrow band methods for variational stereo». In: *German Conference on Pattern Recognition (GCPR)*. 2013.

[13] R. Triebel, J. Stühmer, M. Souiai, and D. Cremers. «Active Online Learning for Interactive Segmentation Using Sparse Gaussian Processes». In: *German Conference on Pattern Recognition (GCPR)*. 2014 (cited on p. 16).

[14] N. Ufer, M. Souiai, and D. Cremers. «Wehrli 2.0: an algorithm for tidying up art». In: *Proc. European Conference on Computer Vision (ECCV), VISART - Where Computer Vision Meets Art - workshop*. Firenze, Italy: Springer, Oct. 2012.

Bibliography

[15] R. Adams and L. Bischof. «Seeded region growing». In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 16.6 (1994), pp. 641–647 (cited on p. 11).

[16] L. Ambrosio, N. Fusco, and D. Pallara. *Functions of bounded variation and free discontinuity problems*. Oxford University Press, Oxford (2000) (cited on p. 67).

[17] L. Ambrosio, N. Fusco, and D. Pallara. *Functions of bounded variation and free discontinuity problems*. Oxford Mathematical Monographs, 2000.

[18] S. Baker, D. Scharstein, J. Lewis, S. Roth, M. J. Black, and R. Szeliski. «A database and evaluation methodology for optical flow». In: *International Journal of Computer Vision (IJCV)* 92.1 (2011), pp. 1–31 (cited on p. 105).

[19] T. Basha, Y. Moses, and N. Kiryati. «Multi-view scene flow estimation: A view centered variational approach». In: *International Journal of Computer Vision (IJCV)* 101.1 (2013), pp. 6–21 (cited on p. 93).

[20] P. Blomgren, T. F. Chan, P. Mulet, C.-K. Wong, et al. «Total variation image restoration: numerical methods and extensions.» In:

[21] A. Bosch, A. Zisserman, and X. Muñoz. «Scene classification via pLSA». In: *Proc. European Conference on Computer Vision (ECCV)*. 2006 (cited on p. 66).

[22] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. «Distributed optimization and statistical learning via the alternating direction method of multipliers». In: *Foundations and trends® in machine learning* 3.1 (2011), pp. 1–122.

[23] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004 (cited on p. 31).

[24] Y. Boykov and V. Kolmogorov. «An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision». In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 26.9 (2004), pp. 1124–1137 (cited on p. 13).

[25] Y. Boykov, O. Veksler, and R. Zabih. «Fast approximate energy minimization via graph cuts». In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23.11 (2001), pp. 1222–1239 (cited on pp. 13, 48).

[26] Y. Boykov, O. Veksler, and R. Zabih. «Fast approximate energy minimization via graph cuts». In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23 (2001), p. 2001 (cited on pp. 13, 19, 21).

[27] J. P. Boyle and R. L. Dykstra. «A method for finding projections onto the intersection of convex sets in Hilbert spaces». In: *Advances in order restricted statistical inference*. Springer, 1986, pp. 28–47 (cited on p. 52).

[28] A. Braides. *Gamma-convergence for Beginners*. Oxford lecture series in mathematics and its applications. Oxford University Press, 2002 (cited on p. 37).

[29] X. Bresson, S. Esedoglu, P. Vandergheynst, J.-P. Thiran, and S. Osher. «Fast global minimization of the active contour/snake model». In: *Jmiv* 28.2 (2007), pp. 151–167 (cited on p. 44).

[30] C. R. Brice and C. L. Fennema. «Scene Analysis using Regions». In: *Artificial Intelligence* 1.3 (1970), pp. 205–226 (cited on p. 11).

[31] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. «High Accuracy Optical Flow Estimation Based on a Theory for Warping». In: *Proc. European Conference on Computer Vision (ECCV)*. Ed. by T. Pajdla and V. Hlavac. Vol. 3024. Prague: Springer, 2004, pp. 25–36.

[32] T. Brox, A. Bruhn, and J. Weickert. «Variational motion segmentation with level sets». In: *Proc. European Conference on Computer Vision (ECCV)*. Lecture Notes in Computer Science. 2006, pp. 471–483 (cited on p. 94).

[33] A. Bruhn, J. Weickert, and C. Schnörr. «Lucas/Kanade meets Horn/Schunck: Combining local and global optic flow methods». In: *International Journal of Computer Vision (IJCV)* 61.3 (2005), pp. 211–231.

[34] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. «A naturalistic open source movie for optical flow evaluation». In: *Proc. European Conference on Computer Vision (ECCV)*. Part IV, LNCS 7577. 2012, pp. 611–625 (cited on p. 101).

[35] E. J. Candès and M. B. Wakin. «An introduction to compressive sampling». In: *Ieee signal processing magazine* 25.2 (2008), pp. 21–30 (cited on p. 51).

[36] J. Cech, J. Sanchez-Riera, and R. Horaud. «Scene flow estimation by growing correspondence seeds». In: *Proc. Int. Conference on Computer Vision and Pattern Recognition (CVPR)*. 2011, pp. 3129–3136.

[37] A. Chambolle, D. Cremers, and T. Pock. *A Convex Approach for Computing Minimal Partitions*. Tech. rep. TR-2008-05. University of Bonn, 2008 (cited on pp. 17, 19, 41, 43).

[38] A. Chambolle, D. Cremers, and T. Pock. «A convex approach to minimal partitions». In: *SIAM Journal on Imaging Sciences* 5.4 (2012), pp. 1113–1158 (cited on pp. 48, 66, 80, 81, 99).

[39] A. Chambolle and T. Pock. «A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging». In: *Jmiv* 40.1 (2011), pp. 120–145 (cited on pp. 23, 33, 35, 70, 79).

[40] A. Chambolle and T. Pock. «A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging». In: *Journal of Mathematical Imaging and Vision (JMIV)* 40.1 (2011), pp. 120–145.

[41] T. Chan, S. Esedoglu, F. Park, and A Yip. «Recent developments in total variation image restoration». In: () (cited on p. 51).

[42] T. F. Chan and J. Shen. «Variational Image Deblurring - A Window into Mathematical Image Processing». In: *Lecture note series, institute for mathematical sciences, national university of singapore* (2004) (cited on p. 99).

[43] T. F. Chan and L. A. Vese. «Active contours without edges». In: *Image processing, IEEE transactions on* 10.2 (2001), pp. 266–277 (cited on p. 12).

[44] G. Charpiat, O. Faugeras, and R. Keriven. «Approximations of shape metrics and application to shape warping and empirical shape statistics». In: *Foundations of Computational Mathematics* 5.1 (2005), pp. 1–58 (cited on p. 20).

[45] Y. Chen, L. Zhu, and A. Yuille. «Active Mask Hierarchies for Object Detection». In: *European conference of computer vision and pattern recognition (eccv)*. 2010, pp. 43–56.

[46] L. D. Cohen. «On active contour models and balloons». In: *CVGIP: Image understanding* 53.2 (1991), pp. 211–218 (cited on p. 20).

[47] D. Cremers, T. Kohlberger, and C. Schnörr. «Shape Statistics in Kernel Space for Variational Image Segmentation». In: *Pattern Recognition* 36.9 (2003), pp. 1929–1943 (cited on p. 20).

[48] D. Cremers and S. Soatto. «Motion Competition: A Variational Approach to Piecewise Parametric Motion Segmentation». In: *International Journal of Computer Vision (IJCV)* 62 (3 2005), pp. 249–265 (cited on p. 94).

[49] P. Das, O. Veksler, V. Zavadsky, and Y. Boykov. «Semiautomatic segmentation with compact shape prior». In: *Image and vision computing* 27.1 (2009), pp. 206–219.

[50] A. Delong, L. Gorelick, O. Veksler, and Y. Boykov. «Minimizing Energies with Hierarchical Costs». English. In: *International Journal of Computer Vision (IJCV)* 100.1 (2012), pp. 38–58 (cited on pp. 21, 22, 25, 59, 62, 65, 67).

[51] A. Delong, A. Osokin, H. N. Isack, and Y. Boykov. «Fast approximate energy minimization with label costs». In: *International journal of computer vision* 96.1 (2012), pp. 1–27 (cited on pp. 25, 43).

[52] J. Diebold, N. Demmel, C. Hazirbas, M. Möller, and D. Cremers. «Interactive Multilabel Segmentation of RGB-D Images». In: *Scale space and variational methods in computer vision (ssvm)*. 2015 (cited on p. 16).

[53] J. Duran, M. Moeller, C. Sbert, and D. Cremers. «Collaborative total variation: a general framework for vectorial TV models». In: *SIAM Journal on Imaging Sciences* 9.1 (2016), pp. 116–151 (cited on p. 51).

[54] J. Eckstein. «Splitting methods for monotone operators with applications to parallel optimization». PhD thesis. Massachusetts Institute of Technology, 1989 (cited on p. 33).

[55] J. Engel, T. Schöps, and D. Cremers. «LSD-SLAM: large-scale direct monocular SLAM». In: *Proc. European Conference on Computer Vision (ECCV)*. 2014 (cited on p. 61).

[56] A. P. Eriksson, C. Olsson, and F. Kahl. «Normalized Cuts Revisited: A Reformulation for Segmentation with Linear Grouping Constraints». In: *Journal of Mathematical Imaging and Vision (JMIV)* 39.1 (2011), pp. 45–61 (cited on p. 83).

[57] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. *Pascal (VOC2010) Results*.

[58] H. Federer. *Geometric measure theory*. Grundlehren der mathematischen Wissenschaften. Springer, 1969 (cited on p. 14).

[59] P. F. Felzenszwalb and D. P. Huttenlocher. «Efficient Graph-Based Image Segmentation». In: *Int. j. comput. vision* 59.2 (Sept. 2004), pp. 167–181.

[60] P. F. Felzenszwalb and O. Veksler. «Tiered scene labeling with dynamic programming». In: *Cvpr*. IEEE. 2010, pp. 3097–3104 (cited on p. 20).

[61] W. Fenchel. «Über konvexe Funktionen mit vorgeschriebenen Niveaumannigfaltigkeiten». In: *Mathematische zeitschrift* 63.1 (1955), pp. 496–506 (cited on p. 27).

[62] M. Feng, J. E. Mitchell, J.-S. Pang, X. Shen, and A. Wächter. «Complementarity formulations of l0-norm optimization problems». In: ().

[63] G. D. Finlayson, M. S. Drew, and C. Lu. «Intrinsic images by entropy minimization». In: *Proc. European Conference on Computer Vision (ECCV)*. Springer, 2004, pp. 582–595 (cited on pp. 48, 77).

[64] G. Gasso, A. Rakotomamonjy, and S. Canu. «Recovering sparse signals with a certain family of nonconvex penalties and DC programming». In: *IEEE Transactions on Signal Processing* 57.12 (2009), pp. 4686–4698 (cited on p. 111).

[65] D. Geman and G. Reynolds. «Constrained restoration and the recovery of discontinuities». In: *Ieee transactions on pattern analysis and machine intelligence* 14.3 (1992), pp. 367–383.

[66] S. Ghadimi and G. Lan. «Accelerated gradient methods for nonconvex nonlinear and stochastic programming». In: *Mathematical Programming* 156.1-2 (2016), pp. 59–99 (cited on p. 51).

[67] G. Gilboa and S. Osher. «Nonlocal Operators with Applications to Image Processing». In: *Multiscale modeling & simulation* 7.3 (2008), pp. 1005–1028 (cited on p. 17).

[68] E. Giusti. *Minimal surfaces and functions of bounded variation / [by] Enrico Giusti ; notes by Graham H. Williams*. English. Australian National University, Canberra, 1977, xi,185p. ; (cited on p. 14).

[69] A. V. Goldberg and R. E. Tarjan. «A new approach to the maximum-flow problem». In: *Journal of the ACM (JACM)* 35.4 (1988), pp. 921–940 (cited on p. 13).

[70] B. Goldluecke and D. Cremers. «Convex Relaxation for Multilabel Problems with Product Label Spaces». In: *Proc. European Conference on Computer Vision (ECCV)*. 2010 (cited on p. 67).

[71] B. Goldluecke and D. Cremers. «Introducing Total Curvature for Image Processing». In: *Proc. Int. Conference on Computer Vision (ICCV)*. 2011.

[72] B. Goldluecke, E. Strekalovskiy, and D. Cremers. «The Natural Total Variation Which Arises from Geometric Measure Theory». In: *SIAM Journal on Imaging Sciences* 5.2 (2012), 537—563 (cited on p. 51).

[73] B. Goldluecke, E. Strekalovskiy, and D. Cremers. «Tight convex relaxations for vector-valued labeling». In: *SIAM Journal on Imaging Sciences* (2013).

[74] L. Gorelick, A. Delong, O. Veksler, and Y. Boykov. «[Recursive MDL via graph cuts: Application to segmentation». In: *Iccv*. 2011, pp. 890–897.

[75] J.-M. Gottfried, J. Fehr, and C. S. Garbe. «Computing range flow from multi-modal Kinect data». In: *Advances in visual computing*. Springer, 2011, pp. 758–767.

[76] D. Greig, B. Porteous, and A. H. Seheult. «Exact maximum a posteriori estimation for binary images». In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1989), pp. 271–279 (cited on p. 13).

[77] T. Gurdan, M. R. Oswald, D. Gurdan, and D. Cremers. «Spatial and Temporal Interpolation of Multi-View Image Sequences». In: *Pattern Recognition - German Conference, (GCPR)*. Vol. 36. Münster, Germany, Sept. 2014 (cited on p. 61).

[78] S. Hadfield and R. Bowden. «Kinecting the dots: Particle based scene flow from depth sensors». In: *Proc. Int. Conference on Computer Vision (ICCV)*. 2011, pp. 2290–2295.

[79] S. Hadfield and R. Bowden. «Scene Particles: Unregularized Particle Based Scene Flow Estimation». In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.3 (2014), pp. 564–576.

[80] C. Hane, C. Zach, A. Cohen, R. Angst, and M. Pollefeys. «Joint 3D scene reconstruction and class segmentation». In: *Proc. Int. Conference on Computer Vision and Pattern Recognition (CVPR)*. 2013, pp. 97–104 (cited on p. 20).

[81] C. Hane, C. Zach, A. Cohen, R. Angst, and M. Pollefeys. «Joint 3D scene reconstruction and class segmentation». In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2013, pp. 97–104 (cited on p. 62).

[82] M. Harris et al. «Optimizing parallel reduction in CUDA». In: *NVIDIA Developer Technology* 2.4 (2007) (cited on p. 43).

[83] E. Herbst, X. Ren, and D. Fox. «RGB-D flow: dense 3-D motion estimation using color and depth». In: *Proc. Int. Conference on Robotics and Automation (ICRA)*. 2013, pp. 2276–2282 (cited on p. 94).

[84] J.-B. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of convex analysis*. Springer Science & Business Media, 2012.

[85] A. J. Hoffman and J. B. Kruskal. «Integral boundary points of convex polyhedra». In: *50 Years of Integer Programming 1958-2008*. Springer, 2010, pp. 49–76 (cited on p. 48).

[86] M. Hornacek, A. Fitzgibbon, and C. Rother. «SphereFlow: 6 DoF Scene Flow from RGB-D Pairs». In: *Proc. Int. Conference on Computer Vision and Pattern Recognition (CVPR)*. 2014, pp. 3526 –3533 (cited on p. 94).

[87] F. Huguet and F. Devernay. «A variational method for scene flow estimation from stereo sequences». In: *Proc. Int. Conference on Computer Vision (ICCV)*. 2007, pp. 1–7 (cited on p. 93).

[88] Institut national de recherche en informatique et en automatique (INRIA) Rhône Alpes. *4d repository*. `http://4drepository.inrialpes.fr/`. Grenoble (cited on pp. 87, 88).

[89] M. Jaimez and J. Gonzalez-Jimenez. «Fast Visual Odometry for 3-D Range Sensors». In: *IEEE Transactions on Robotics* 31.4 (2015), pp. 809–822 (cited on p. 97).

[90] C. Jordan. *Sur la serie de fourier*. Comptes rendus hebdomadaires des seances de l'Academie des sciences (in French) 92: 228Ð230, JFM 13.0184.01, 1881 (cited on p. 13).

[91] J. Kappes and C. Schnörr. «MAP-inference for highly-connected graphs with DC-programming». In: *Joint Pattern Recognition Symposium*. Springer. 2008, pp. 1–10 (cited on p. 50).

[92] J. Kappes and C. Schnörr. «MAP-inference for highly-connected graphs with DC-programming». In: *Joint Pattern Recognition Symposium*. Springer, 2008, pp. 1–10 (cited on p. 78).

[93] M. Kass, A. Witkin, and D. Terzopoulos. «Snakes: Active contour models». In: *International Journal of Computer Vision (IJCV)* 1.4 (1988), pp. 321–331 (cited on pp. 12, 20).

[94] C. Kerl, J. Sturm, and D. Cremers. «Robust Odometry Estimation for RGB-D Cameras». In: *Proc. Int. Conference on Robotics and Automation (ICRA)*. 2013 (cited on pp. 61, 97).

[95] M. Klodt, T. Schoenemann, K. Kolev, M. Schikora, and D. Cremers. «An Experimental Comparison of Discrete and Continuous Shape Optimization Methods». In: *Proc. European Conference on Computer Vision (ECCV)*. Marseille, France, 2008 (cited on pp. 13, 59, 62, 77).

[96] M. Klodt, F. Steinbruecker, and D. Cremers. «Moment Constraints in Convex Optimization for Segmentation and Tracking». In: *Advanced topics in computer vision*. Springer, 2013 (cited on pp. 20, 77, 83).

[97] M. Klodt, J. Sturm, and D. Cremers. «Scale-aware object tracking with convex shape constraints on rgb-d images». In: *German Conference on Pattern Recognition (GCPR)*. Saarbrücken, Germany, 2013 (cited on p. 20).

[98] P. Kohli, L. Ladicky, and P. H. S. Torr. «Robust higher order potentials for enforcing label consistency». In: *International Journal of Computer Vision (IJCV)* 82.3 (2009), pp. 302–324.

[99] N. Komodakis, N. Paragios, and G. Tziritas. «MRF Optimization via Dual Decomposition: Message-Passing Revisited». In: *Iccv*. 2007, pp. 1–8.

[100] N. Komodakis and G. Tziritas. «Approximate labeling via graph cuts based on linear programming». In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29.8 (2007), pp. 1436–1453 (cited on p. 13).

[101] N. Komodakis and G. Tziritas. «Approximate labeling via graph cuts based on linear programming». In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29.8 (2007), pp. 1436–1453.

[102] T. P. Kristian Bredies Karl Kunisch. «Total Generalized Variation». In: *Preprint* (2009).

[103] L. Ladicky, C. Russell, P. Kohli, and P. Torr. «Inference Methods for CRFs with Co-occurrence Statistics». English. In: *International Journal of Computer Vision (IJCV)* 103.2 (2013), pp. 213–225.

[104] L. Ladicky, C. Russell, P. Kohli, and P. H. S. Torr. «Graph Cut Based Inference with Co-occurrence Statistics». In: *Proc. European Conference on Computer Vision (ECCV)*. 2010, pp. 239–253 (cited on pp. 21, 25, 59, 62, 66, 71–73).

[105] K. Lange. «The MM algorithm». In: *Optimization*. Springer, 2013, pp. 185–219 (cited on p. 50).

[106] Y. G. Leclerc. «Region growing using the MDL principle». In: *Proc. darpa image underst. workshop*. 1990, pp. 720–726 (cited on pp. 21, 62, 65).

[107] J. Lellmann, J. Kappes, J. Yuan, F. Becker, and C. Schnörr. *Convex Multi-Class Image Labeling by Simplex-Constrained Total Variation.* Tech. rep. Heidelberg University, 2008 (cited on p. 66).

[108] J. Lellmann, J. H. Kappes, J. Yuan, F. Becker, and C. Schnörr. «Convex Multi-Class Image Labeling by Simplex-Constrained Total Variation». In: *Ssvm 2009.* Ed. by X.-C. Tai, K. Mórken, M. Lysaker, and K.-A. Lie. Vol. 5567. LNCS. Springer, 2009, pp. 150–162.

[109] J. Lellmann, F. Lenzen, and C. Schnörr. «Optimality bounds for a variational relaxation of the image partitioning problem». In: *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition.* Springer. 2011, pp. 132–146 (cited on p. 19).

[110] J. Lellmann, F. Lenzen, and C. Schnörr. «Optimality Bounds for a Variational Relaxation of the Image Partitioning Problem». In: *Proceedings of the International Conference on Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR).* 2011, pp. 132–146 (cited on p. 81).

[111] J. Lellmann, F. Lenzen, and C. Schnörr. «Optimality Bounds for a Variational Relaxation of the Image Partitioning Problem». In: *Journal of Mathematical Imaging and Vision (JMIV)* 47.3 (2013), pp. 239–257 (cited on p. 81).

[112] A. Letouzey, B. Petit, and E. Boyer. «Scene Flow from Depth and Color Images». In: *Proc. British Machine Vision Conference (BMVC).* 2011, pp. 46.1–46.11.

[113] L.-J. Li, H. Su, Y. Lim, and L. Fei-Fei. «Objects as attributes for scene classification». In: *eccv, International Workshop on Parts and Attributes.* 2010 (cited on p. 66).

[114] L.-J. Li, H. Su, E. P. Xing, and L. Fei-Fei. «Object Bank: A High-Level Image Representation for Scene Classification and Semantic Feature Sparsification». In: *Proc. Neural Information Processing Systems.* 2010 (cited on p. 66).

[115] Y. Lim, K. Jung, and P. Kohli. «Energy Minimization under Constraints on Label Counts». In: *Proc. European Conference on Computer Vision (ECCV).* 2010, pp. 535–551 (cited on p. 83).

[116] X. Liu, O. Veksler, and J. Samarabandu. «Order-preserving moves for graph-cut-based optimization». In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32.7 (2010), pp. 1182–1196 (cited on p. 20).

[117] J. Long, E. Shelhamer, and T. Darrell. «Fully convolutional networks for semantic segmentation». In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 2015, pp. 3431–3440.

[118] W. E. Lorensen and H. E. Cline. «Marching cubes: A high resolution 3D surface construction algorithm». In: *Siggraph comput. graph.* 21 (4 1987), pp. 163–169 (cited on p. 87).

[119] L. Lovász. «On the ratio of optimal integral and fractional covers». In: *Discrete mathematics* 13.4 (1975), pp. 383–390 (cited on p. 15).

[120] B. D. Lucas and T. Kanade. «An iterative image registration technique with an application to stereo vision». In: *Proc.7th International Joint Conference on Artificial Intelligence.* Vancouver, 1981, pp. 674–679.

[121] B. D. Lucas, T. Kanade, et al. «An iterative image registration technique with an application to stereo vision.» In: 1981 (cited on p. 61).

[122] Z. Ma, K. He, Y. Wei, J. Sun, and E. Wu. «Constant time weighted median filtering for stereo matching and beyond». In: *Proc. Int. Conference on Computer Vision (ICCV)*. 2013, pp. 49–56.

[123] D. J. MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003 (cited on p. 20).

[124] J. MacQueen. «Some methods for classification and analysis of multivariate observations». In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*. Berkeley, Calif.: University of California Press, 1967, pp. 281–297.

[125] J. Malik, S. Belongie, T. Leung, and J. Shi. «Contour and Texture Analysis for Image Segmentation». In: *Int. j. comput. vision* 43.1 (June 2001), pp. 7–27.

[126] C. Michelot. «A finite algorithm for finding the projection of a point onto the canonical simplex of R^n». In: *Journal of optimization theory and applications* 50.1 (1986) (cited on p. 42).

[127] G. J. Minty et al. «Monotone (nonlinear) operators in Hilbert space». In: *Duke mathematical journal* 29.3 (1962), pp. 341–346 (cited on p. 32).

[128] G. J. Minty et al. «Monotone (nonlinear) operators in Hilbert space». In: *Duke mathematical journal* 29.3 (1962), pp. 341–346.

[129] T. Möllenhoff, E. Strekalovskiy, M. Möller, and D. Cremers. «Low Rank Priors for Color Image Regularization». In: *Proceedings of the International Conference on Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR)*. 2015 (cited on p. 78).

[130] J.-J. Moreau. «Proximité et dualité dans un espace hilbertien». In: *Bulletin de la Société mathématique de France* 93 (1965), pp. 273–299 (cited on p. 27).

[131] D. Mumford and J. Shah. «Optimal approximations by piecewise smooth functions and associated variational problems». In: *Comm. pure appl. math.* 42.5 (1989), pp. 577–685 (cited on pp. 20, 44).

[132] L. Najman and M. Schmitt. «Watershed of a continuous function». In: *Signal Processing* 38.1 (1994), pp. 99–112.

[133] G. L. Nemhauser and L. A. Wolsey. «Integer programming and combinatorial optimization». In: *Wiley, Chichester. GL Nemhauser, MWP Savelsbergh, GS Sigismondi (1992). Constraint Classification for Mixed Integer Programming Formulations. COAL Bulletin* 20 (1988), pp. 8–12 (cited on p. 48).

[134] M. Niethammer and C. Zach. «Segmentation with area constraints». In: *Medical Image Analysis* 17.1 (2013), pp. 101–112 (cited on p. 83).

[135] C. Nieuwenhuis and D. Cremers. «Spatially Varying Color Distributions for Interactive Multi-Label Segmentation». In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.5 (2013), pp. 1234–1247 (cited on p. 16).

[136] C. Nieuwenhuis, E. Strekalovskiy, and D. Cremers. «Proportion Priors for Image Sequence Segmentation». In: *Proc. Int. Conference on Computer Vision (ICCV)*. Sydney, Australia, 2013 (cited on pp. 20, 77).

[137] C. Nieuwenhuis, E. Toeppe, and D. Cremers. «A Survey and Comparison of Discrete and Continuous Multilabel Segmentation Approaches». In: *International Journal of Computer Vision (IJCV)* (2013) (cited on pp. 59, 66, 82).

[138] M. Nikolova, S. Esedoglu, and T. F. Chan. «Algorithms for finding global minimizers of image segmentation and denoising models». In: *SIAM Journal on Applied Mathematics* 66.5 (2006), pp. 1632–1648 (cited on pp. 14, 15, 20, 48, 77, 83, 99).

[139] P. Ochs, J. Malik, and T. Brox. «Segmentation of moving objects by long term video analysis». In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.6 (2014). Preprint, pp. 1187 –1200.

[140] P. Ochs, A. Dosovitskiy, T. Brox, and T. Pock. «An iterated l1 algorithm for non-smooth non-convex optimization in computer vision». In: *Proc. Int. Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. 2013, pp. 1759–1766 (cited on p. 78).

[141] S. Osher and J. A. Sethian. «Fronts propagating with curvature-dependent speed: algorithms based on hamilton-jacobi formulations». In: *Journal of computational physics* 79.1 (1988), pp. 12–49 (cited on p. 12).

[142] M. R. Oswald and D. Cremers. «A Convex Relaxation Approach to Space Time Multi-view 3D Reconstruction». In: *ICCV Workshop on Dynamic Shape Capture and Analysis (4DMOD)*. 2013 (cited on pp. 9, 86, 87).

[143] M. R. Oswald, E. Töppe, and D. Cremers. «Fast and Globally Optimal Single View Reconstruction of Curved Objects». In: *Proc. Int. Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2012, pp. 534–541 (cited on p. 83).

[144] K. Pandit, C. Kirchner, J. Schmitt, and R. Steinmetz. *Optimization of the min-plus convolution computation under Network Calculus constraints*. Tech. rep. (cited on p. 29).

[145] N. Parikh, S. P. Boyd, et al. «Proximal algorithms.» In: () (cited on p. 53).

[146] T. Pock and A. Chambolle. «[diagonal preconditioning for first order primal-dual algorithms in convex optimization». In: (cited on p. 70).

[147] T. Pock, D. Cremers, H. Bischof, and A. Chambolle. «An Algorithm for Minimizing the Piecewise Smooth Mumford-Shah Functional». In: *Proc. Int. Conference on Computer Vision (ICCV)*. 2009 (cited on pp. 23, 33, 35, 48).

[148] T. Pock, D. Cremers, H. Bischof, and A. Chambolle. «An Algorithm for Minimizing the Piecewise Smooth Mumford-Shah Functional». In: *Proc. Int. Conference on Computer Vision (ICCV)*. Kyoto, Japan, 2009.

[149] T. Pock, D. Cremers, H. Bischof, and A. Chambolle. «Global Solutions of Variational Models with Convex Regularization». In: *SIAM Journal on Imaging Sciences* 3.4 (2010), pp. 1122–1145 (cited on p. 48).

[150] T. Pock and A. Chambolle. «Diagonal preconditioning for first order primal-dual algorithms in convex optimization». In: *Proc. Int. Conference on Computer Vision (ICCV)*. 2011 (cited on pp. 35, 98).

[151] T. Pock, T. Schoenemann, G. Graber, H. Bischof, and D. Cremers. «A Convex Formulation of Continuous Multi-label Problems». In: *Proc. European Conference on Computer Vision (ECCV)*. 2008, pp. 792–805 (cited on p. 76).

[152] R. B. Potts. «Some generalized order-disorder transformations». In: *Mathematical proceedings of the cambridge philosophical society*. Vol. 48. 01. Cambridge Univ Press. 1952, pp. 106–109.

[153] J. Quiroga, T. Brox, F. Devernay, and J. Crowley. «Dense semi-rigid scene flow estimation from RGBD images». In: *Proc. European Conference on Computer Vision (ECCV)*. 2014, pp. 567–582 (cited on pp. 94, 102, 104).

[154] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. «Objects in context». In: *Proc. Int. Conference on Computer Vision (ICCV)*. IEEE. 2007, pp. 1–8 (cited on p. 21).

[155] C. Reinbacher, T. Pock, C. Bauer, and H. Bischof. «Variational Segmentation of Elongated Volumetric Structures». In: *Proc. Int. Conference on Computer Vision and Pattern Recognition (CVPR)*. 2010 (cited on pp. 77, 83).

[156] R. T. Rockafellar. «Monotone operators and the proximal point algorithm». In: *SIAM journal on control and optimization* 14.5 (1976), pp. 877–898 (cited on pp. 31, 32).

[157] R. T. Rockafellar. *Convex analysis*. Princeton university press, 2015 (cited on pp. 27, 51).

[158] E. Rodola, S. R. Bulò, and D. Cremers. «Robust region detection via consensus segmentation of deformable shapes». In: *Computer Graphics Forum*. Vol. 33. 5. Wiley Online Library. 2014, pp. 97–106 (cited on p. 9).

[159] J. B. Roerdink and A. Meijster. «The watershed transform: Definitions, algorithms and parallelization strategies». In: *Fundamenta informaticae* 41.1, 2 (2000), pp. 187–228 (cited on p. 11).

[160] A. Roussos, C. Russell, R. Garg, and L. de Agapito. «Dense multibody motion estimation and reconstruction from a handheld camera». In: *IEEE Intl. Symp. on Mixed and Augmented Reality (ISMAR)*. 2012, pp. 31 –40 (cited on p. 94).

[161] L. I. Rudin, S. Osher, and E. Fatemi. «Nonlinear Total Variation Based Noise Removal Algorithms». In: *Physica d* 60 (1992), pp. 259–268 (cited on pp. 13, 51, 98, 99).

[162] B. C. Russell, W. T. Freeman, A. A. Efros, J. Sivic, and A. Zisserman. «Using Multiple Segmentations to Discover Objects and their Extent in Image Collections». In: *Proc. Int. Conference on Computer Vision and Pattern Recognition (CVPR)*. Washington, DC, USA: IEEE Computer Society, 2006, pp. 1605–1614.

[163] T. Schoenemann and D. Cremers. «Globally Optimal Image Segmentation with an Elastic Shape Prior». In: *Proc. Int. Conference on Computer Vision (ICCV)*. Rio de Janeiro, Brazil, Oct. 2007 (cited on p. 20).

[164] A. Schrijver. *Combinatorial optimization: polyhedra and efficiency*. Vol. 24. Springer Science & Business Media, 2002 (cited on p. 48).

[165] J. Shotton, J. M. Winn, C. Rother, and A. Criminisi. «*TextonBoost*: Joint Appearance, Shape and Context Modeling for Multi-class Object Recognition and Segmentation». In: *Proc. European Conference on Computer Vision (ECCV)*. 2006, pp. 1–15 (cited on p. 16).

[166] D. Shulman and J.-Y. Herve. «Regularization of discontinuous flow fields». In: *Visual motion, 1989., proceedings. workshop on*. IEEE. 1989, pp. 81–86 (cited on p. 13).

[167] M. Souiai. «Newton Methods for Total Variation Minimization». MA thesis. Germany: Computer Vision Group, TU Munich, 2010.

[168] H. Spies, B. Jähne, and J. L. Barron. «Range flow estimation». In: *Computer Vision and Image Understanding* 85.3 (2002), pp. 209–231.

[169] F. Steinbruecker, T. Pock, and D. Cremers. «Large Displacement Optical Flow Computation without Warping». In: *IEEE International Conference on Computer Vision (iccv)*. Kyoto, Japan, 2009 (cited on p. 50).

[170] H. Steinhaus. «Sur la division des corp materiels en parties». In: *Bull. Acad. Polon. Sci* 1 (1956), pp. 801–804.

[171] P. Strandmark, F. Kahl, and N. C. Overgaard. «Optimizing Parametric Total Variation Models». In: *Proc. Int. Conference on Computer Vision (ICCV)*. 2009.

[172] E. Strekalovskiy and D. Cremers. «Generalized Ordering Constraints for Multilabel Optimization». In: *Proc. Int. Conference on Computer Vision (ICCV)*. 2011 (cited on p. 20).

[173] E. Strekalovskiy and D. Cremers. «Real-Time Minimization of the Piecewise Smooth Mumford-Shah Functional». In: *Proc. European Conference on Computer Vision (ECCV)*. 2014, pp. 127–141 (cited on pp. 50, 54, 55).

[174] E. Strekalovskiy, B. Goldluecke, and D. Cremers. «Tight Convex Relaxations for Vector-Valued Labeling Problems». In: *Proc. Int. Conference on Computer Vision (ICCV)*. 2011 (cited on p. 67).

[175] J. Stückler and S. Behnke. «Efficient Dense Rigid-Body Motion Segmentation and Estimation in RGB-D Video». In: *International Journal of Computer Vision (IJCV)* 113.3 (2015), pp. 233–245 (cited on p. 94).

[176] D. Sun, E. B. Sudderth, and H. Pfister. «Layered RGBD Scene Flow Estimation». In: *Proc. Int. Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015 (cited on pp. 92, 94, 104).

[177] W.-y. Sun, R. J. Sampaio, and M. Candido. «Proximal point algorithm for minimization of DC function». In: *Journal of computational Mathematics* (2003), pp. 451–462.

[178] M. Sussman, P. Smereka, and S. Osher. «A level set approach for computing solutions to incompressible two-phase flow». In: *Journal of Computational physics* 114.1 (1994), pp. 146–159 (cited on p. 12).

[179] R. Takapoui and H. Javadi. «Preconditioning via diagonal scaling». In: (2014) (cited on p. 35).

[180] M. Tang, I. Ben Ayed, and Y. Boykov. «Pseudo-bound optimization for binary energies». English. In: *Proc. European Conference on Computer Vision (ECCV)*. Ed. by D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars. Vol. 8693. Lecture Notes in Computer Science. Springer International Publishing, 2014, pp. 691–707 (cited on pp. 48, 77).

[181] P. D. Tao. «Convex analysis approach to dc programming: Theory, algorithms and applications». In: *Acta Mathematica Vietnamica* 22.1 (1997), pp. 289–355 (cited on p. 79).

[182] P. D. Tao and E. B. Souad. «Algorithms for solving a class of nonconvex optimization problems. Methods of subgradients». In: *North-Holland Mathematics Studies* 129 (1986), pp. 249–271 (cited on pp. 49, 50, 77, 78).

[183] A. N. Tikhonov. «On the stability of inverse problems». In: *Dokl. Akad. Nauk SSSR.* Vol. 39. 5. 1943, pp. 195–198.

[184] J. Toland. «Duality in nonconvex optimization». In: *Journal of Mathematical Analysis and Applications* 66.2 (1978), pp. 399–415 (cited on p. 50).

[185] J. Toland. «Duality in nonconvex optimization». In: *Journal of Mathematical Analysis and Applications* 66.2 (1978), pp. 399–415 (cited on p. 78).

[186] E. Töppe, M. R. Oswald, D. Cremers, and C. Rother. «Image-based 3D Modeling via Cheeger Sets.» In: *Proc. Asian Conference on Computer Vision (ACCV).* Nov. 2010 (cited on pp. 77, 83, 84).

[187] A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin. «Context-based vision system for place and object recognition». In: *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on.* IEEE. 2003, pp. 273–280 (cited on p. 22).

[188] M. Unger, T. Pock, M. Werlberger, and H. Bischof. «A Convex Approach for Variational Super-Resolution». In: *Proceedings German Association for Pattern Recognition (DAGM).* Vol. 6376. LNCS. Heidelberg: Springer, 2010, pp. 313–322 (cited on p. 61).

[189] M. Unger, M. Werlberger, T. Pock, and H. Bischof. «Joint Motion Estimation and Segmentation of Complex Scenes with Label Costs and Occlusion Modeling». In: *Proc. Int. Conference on Computer Vision and Pattern Recognition (CVPR).* 2012 (cited on p. 94).

[190] S. Vedula, S. Baker, P. Rander, R. Collins, and T. Kanade. «Three-dimensional scene flow». In: *Proc. Int. Conference on Computer Vision (ICCV).* Vol. 2. 1999, pp. 722–729 (cited on p. 93).

[191] O. Veksler. «Star shape prior for graph-cut image segmentation». In: *Computer vision–eccv 2008.* Springer, 2008, pp. 454–467.

[192] C. Vogel, K. Schindler, and S. Roth. «3D scene flow estimation with a rigid motion prior». In: *Proc. Int. Conference on Computer Vision (ICCV).* 2011, pp. 1291–1298 (cited on p. 93).

[193] C. Vogel, K. Schindler, and S. Roth. «Piecewise Rigid Scene Flow». In: *Proc. Int. Conference on Computer Vision (ICCV).* 2013, pp. 1377–1384 (cited on pp. 92, 93).

[194] M. J. Wainwright, T. S. Jaakkola, and A. S. Willsky. «MAP estimation via agreement on trees: message-passing and linear programming». In: *Information theory, ieee transactions on* 51.11 (2005), pp. 3697–3717 (cited on p. 13).

[195] S. Walukiewicz. *Integer programming.* Mathematics and its applications East European series. Revised translation from the Polish original Programowanie dyskretne, published in 1986–T.p. verso. Dordrecht, Boston: Kluwer Academic Publishers Warszawa, 1991 (cited on p. 47).

[196] S. Weber. «Discrete Tomography by Convex-Concave Regularization using Linear and Quadratic Optimization». PhD thesis. Ruprecht-Karls-Universität Heidelberg, Germany, 2009 (cited on p. 79).

[197] A. Wedel, T. Brox, T. Vaudrey, C. Rabe, U. Franke, and D. Cremers. «Stereoscopic Scene Flow Computation for 3D Motion Understanding». In: *International Journal of Computer Vision (IJCV)* 95.1 (2011), pp. 29–51 (cited on pp. 61, 93).

[198] A. Wedel, T. Pock, C. Zach, H. Bischof, and D. Cremers. «An improved algorithm for TV-L1 optical flow». In: *Statistical and Geometrical Approaches to Visual Motion Analysis*. Springer, 2009, pp. 23–45.

[199] M. Werlberger, M. Unger, T. Pock, and H. Bischof. «Efficient Minimization of the Non-Local Potts Model». In: *Icssvmcv*. 2011 (cited on pp. 18, 41, 44, 67).

[200] T. Werner. «A linear programming approach to max-sum problem: a review». In: *IEEE transactions on pattern analysis and machine intelligence* 29.7 (2007), pp. 1165–1179 (cited on p. 48).

[201] L. Xu, C. Lu, Y. Xu, and J. Jia. «Image smoothing via L 0 gradient minimization». In: *ACM Transactions on Graphics (TOG)*. Vol. 30. 6. ACM. 2011, p. 174 (cited on pp. 50, 54, 55).

[202] J. Yao, S. Fidler, and R. Urtasun. «Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation». In: *Proc. Int. Conference on Computer Vision and Pattern Recognition (CVPR)*. 2012 (cited on p. 66).

[203] J. Yuan and Y. Boykov. «TV-Based Multi-Label Image Segmentation with Label Cost Prior». In: (cited on p. 25).

[204] J. Yuan and Y. Boykov. «TV-Based Multi-Label Image Segmentation with Label Cost Prior». In: *Bmvc*. 2010, pp. 1–12 (cited on pp. 21, 43, 62, 65, 67–69).

[205] A. L. Yuille and A. Rangarajan. «The Concave-Convex Procedure». In: *Neural Computation* 15.4 (2003), pp. 915–936 (cited on pp. 50, 78).

[206] C. Zach, D. Gallup, J.-M Frahm, and M. Niethammer. «Fast Global Labeling for Real-Time Stereo Using Multiple Plane Sweeps». In: *Vision, modeling and visualization*. 2008, pp. 243–252 (cited on pp. 17, 41, 42, 66, 80, 81).

[207] G. Zhang, J. Jia, and H. Bao. «Simultaneous Multi-Body Stereo and Segmentation». In: *Proc. Int. Conference on Computer Vision (ICCV)*. 2011, pp. 826 –833 (cited on p. 94).

[208] Y. Zhang and C. Kambhamettu. «On 3D scene flow and structure estimation». In: *Proc. Int. Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 2. 2001, pp. 778–785 (cited on p. 93).

[209] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr. «Conditional Random Fields as Recurrent Neural Networks». In: *Proc. Int. Conference on Computer Vision (ICCV)* (2015) (cited on p. 16).

[210] L. Zhu, Y. Chen, A. Yuille, and W. Freeman. «Latent Hierarchical Structural Learning for Object Detection». In: *International conference for computer vision and pattern recognition (cvpr)*. 2010.

[211] S. C. Zhu and A. Yuille. «Region Competition: Unifying Snakes, Region Growing, and Bayes/MDL for Multi-band Image Segmentation». In: *PAMI* 18 (1996), pp. 884–900 (cited on pp. 21, 62, 65).