Convex Variational Image Restoration with Histogram Priors

Paul Swoboda and Christoph Schnörr

July 18, 2013

Abstract

We present a novel variational approach to image restoration (e.g., denoising, inpainting, labeling) that enables to complement established variational approaches with a histogram-based prior enforcing closeness of the solution to some given empirical measure. By minimizing a single objective function, the approach utilizes simultaneously two quite different sources of information for restoration: spatial context in terms of some smoothness prior and non-spatial statistics in terms of the novel prior utilizing the Wasserstein distance between probability measures. We study the combination of the functional lifting technique with two different relaxations of the histogram prior and derive a jointly convex variational approach. Mathematical equivalence of both relaxations is established and cases where optimality holds are discussed. Additionally, we present an efficient algorithmic scheme for the numerical treatment of the presented model. Experiments using the basic total-variation based denoising approach as a case study demonstrate our novel regularization approach.

1 Introduction

A broad range of powerful variational approaches to low-level image analysis tasks exist, like image denoising, image inpainting or image labeling [12, 13]. It is not straightforward however to incorporate *directly* into the restoration process statistical prior knowledge about the image class at hand. Particularly, handling global statistics as part of a single *convex* variational approach has not been considered so far.

In the present paper, we introduce a class of variational approaches of the form

$$\inf_u F(u) + \lambda R(u) + \nu W(\mu^u, \mu^0), \tag{1.1}$$

where $F(u) + \lambda R(u)$ is any energy functional consisting of a data fidelity term $F(u)$ and a regularization term $R(u)$, $W(\mu^u, \mu^0)$ denotes the histogram prior in terms of the Wasserstein distance between the histogram corresponding to the minimizing function u to be determined and some given histogram μ^0 and $\lambda > 0$ and $\nu > 0$ are parameters weighing the influence of each term. We require $R(u)$

1
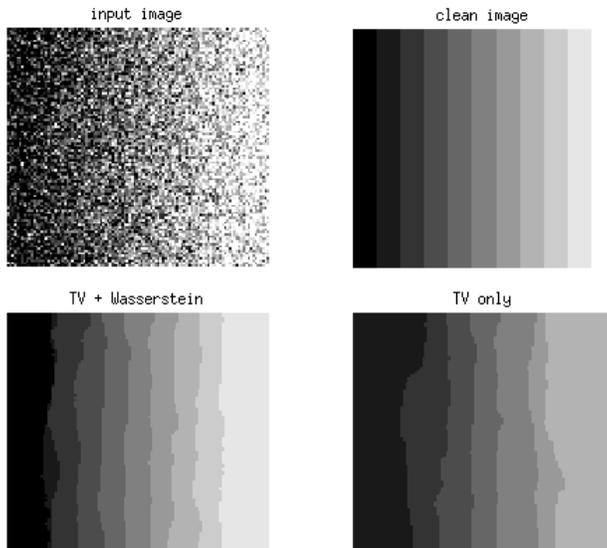
Figure 1: Denoising experiment of a noisy image (upper row, left side) taking into account statistical prior information through convex optimization (lower row, left side) infers the correct image structure and outperforms hand-tuned established variational restoration (lower row, right side). Enforcing global image statistics to be similar to those of the clean image (upper row, right side) gives our approach an advantage over methods not taking such information into account.

to be convex. As a case study, we adopt for $R(u) = \text{TV}(u)$, the Total Variation, see [2], and $F(u) = \int_\Omega f(u(x), x)dx$, where f can also be a nonconvex function. The basic ROF denoising approach of [22] is included in this approach with $f(u(x), x) = (u(x) - u_0(x))^2$, where u_0 is the image to be denoised.

Note that minimizing the second term $R(u)$ in (1.1) entails spatial regularization whereas the third Wasserstein term utilizes statistical information that is not spatially indexed in any way. As an illustration, consider the academical example in figure 1. Knowing the grayvalue distribution of the original image helps us in regularizing the noisy input image. We tackle the corresponding main difficulty in two different, mathematically plausible ways: by convex relaxations of (1.1) in order to obtain a computationally tractable approach. Comparing these two relaxations – one may be tighter than the other one – reveals however mathematical equivalence. Preliminary numerical experiments demonstrate that the relaxation seems to be tight enough so as to bias effectively variational restoration towards given statistical prior information.

2 Prior Work, Contribution

2.1 Related Work

Image regularization by variational methods is a powerful and commonly used tool for denoising, inpainting, labeling and many other applications. As a case study in connection with (1.1), we consider one of the most widely used approaches for denoising, namely the Rudin, Osher and Fatemi (ROF) model from [22]:

$$\min_{u \in \mathrm{BV}(\Omega,[0,1])} \|u - u_0\|^2 + \lambda \mathrm{TV}(u), \tag{2.1}$$

where u_0 is the input image, TV denotes the Total Variation and $\mathrm{BV}(\Omega, [0,1])$ is the space of functions of bounded variation with domain $\Omega \subset \mathbb{R}^d$ and values in $[0,1]$. The minimization problem (2.1) is convex and can be solved to a global optimum efficiently by various first-order proximal splitting algorithms even for large problem sizes, e.g. by Primal-Dual methods [5] or other proximal minimization algorithms for nonsmooth convex optimization [3, 7, 19].

We can also use more general data terms instead of the quadratic term in (2.1). For example in [15] it is shown how the data term can be replaced by a continuous but possibly non-convex function $\int_\Omega f(u(x), x) dx$. Still this data function is local and does not take into account global statistics.

In the case that some prior knowledge is encoded as a histogram, the Wasserstein distance and the associated Optimal Transport are a suitable choice for penalizing deviance from prior knowledge. More generally the Wasserstein distance can be used as a distance on histograms over arbitrary metricized spaces.

Regarding the Wasserstein distance and the theory of Optimal Transport we refer to the in-depth treatise [23]. Optimal Transport is well-known as Earth Mover's distance in image processing and computer vision [21] and has been used for content-based image retrieval. Further recent applications include [6, 14] in connection with segmentation and [8] for texture synthesis.

The authors of [17] propose an approach to contrast and color modification. Given a prior model of how the color or grayvalues are distributed in an image, the authors propose a variational formulation for modifying the given image so that these statistical constraints are met in a spatially regular way. While their algorithm is fast, high runtime performance is achieved by minimizing a non-convex approximation of their original energy. In contrast, we directly minimize a convex relaxation of the original energy, hence we may hope to obtain lower energies and not to get stuck in local minima.

Our variational approach employing the Wasserstein distance as a histogram-based prior through *convex* relaxation appears to be novel.

2.2 Contribution

We present

- a variational model with a histogram-based prior for image restoration (Section 3),

3

- two convex relaxations of the original problem together with discussions of cases where optimality holds (Sections 4 and 5),

- a proof of equivalence for the two presented relaxations (Section 6),

- an efficient numerical implementation of the proposed variational model (Section 7),

- experimental validation of the proposed approach (Section 8).

3 Problem and Mathematical Background

We introduce the original non-convex model, consider different ways to write the Wasserstein distance and introduce the functional lifting technique for rewriting the resulting optimization problem to show well-posedness and to make it amenable for global optimization.

3.1 Problem Statement

For an image domain $\Omega \subset \mathbb{R}^2$, e.g. $\Omega = [0,1]^2$ and $u : \Omega \to [0,1]$, consider the normalized pushforward $\mathcal{L}_{|\Omega}$ of the Lebesgue measure $\mathcal{L}$ restricted to Ω by u:

$$\mu^u(A) = \frac{1}{\mathcal{L}(\Omega)} \left(u_* \mathcal{L} \right)(A) = \frac{1}{\mathcal{L}(\Omega)} \mathcal{L}(u^{-1}(A)) \quad \forall A \subset [0,1] \text{ measurable}. \quad (3.1)$$

We will use the notation $|B| := \mathcal{L}(B)$ for simplicity. In other words, μ^u is the grayvalue histogram of the image u. We would like to minimize the energy function

$$\min_{u \in \mathrm{BV}(\Omega,[0,1])} E(u) = \int_\Omega f(u(x),x)dx + \lambda \mathrm{TV}(u) + \nu W(\mu^u, \mu^0). \quad (3.2)$$

$\mathrm{TV}(u)$ is the Total Variation

$$\mathrm{TV}(u) = \sup \left\{ \int_\Omega u(x) \operatorname{div} \phi(x) dx \ : \ \phi \in C_c^1(\Omega, \mathbb{R}^2), \|\phi\|_\infty \leq 1 \right\}, \quad (3.3)$$

where $C_c^1(\Omega, \mathbb{R}^2)$ is the space of continuously differentiable functions with compact support in Ω and values in $\mathbb{R}^2$, see [2] for more details. $f : [0,1] \times \Omega \to \mathbb{R}$ is a continuous fidelity function, and W is the Wasserstein distance

$$W(\mu, \tilde{\mu}) = \inf_{\pi \in \Pi(\mu, \tilde{\mu})} \int_{[0,1] \times [0,1]} c(\gamma_1, \gamma_2) \ d\pi(\gamma_1, \gamma_2). \quad (3.4)$$

$c : [0,1] \times [0,1] \to \mathbb{R}$ is the cost function for the Wasserstein distance, for example $c(\gamma_1, \gamma_2) = |\gamma_1 - \gamma_2|^p$ with $p \geq 1$. The space of transport plans is

$$\Pi(\mu, \tilde{\mu}) = \{ \pi \in \mathcal{P}([0,1] \times [0,1]) \ : \ \begin{matrix} \pi(A \times [0,1]) = \mu(A) \\ \pi([0,1] \times B) = \tilde{\mu}(B) \end{matrix} \quad \forall A, B \text{ measurable}\}, \quad (3.5)$$

where $\mathcal{P}([0,1] \times [0,1])$ is the space of all probability measures defined on the Borel-σ-Algebra over $[0,1] \times [0,1]$. If c is lower semicontinuous and there exist upper semicontinuous functions $a, b \in L_1([0,1])$ such that $c(\gamma_1, \gamma_2) \geq a(\gamma_1) + b(\gamma_2)$, then by Theorem 4.1 in [23] there exists a measure which minimizes (3.4) and which is called the optimal transport plan. The optimization problem (3.4) is linear in both constraints and objective and therefore convex. Note however that energy (3.2) is not convex.

By minimizing (3.2) we obtain a solution u which remains faithful to the data by the fidelity term f, is spatially coherent by the Total Variation term and has global grayvalue statistics similar to μ^0 by the Wasserstein term.

Remark 1. In the case of two labels, which means restricting the function u in (3.2) to have values $u(x) \in \{0, 1\}$, our model reduces to foreground/background segmentation and the Wasserstein term can be interpreted as a prior favoring a prespecified size of the foreground area.

3.2 The Wasserstein Distance and its Dual

We reformulate energy (3.2) by introducing another way to obtain the Wasserstein distance. Assume the cost $c : [0,1] \times [0,1] \to \mathbb{R}$ is lower semicontinuous such that

$$c(\gamma_1, \gamma_2) \geq a(\gamma_1) + b(\gamma_2) \quad \forall x, y \in [0,1] \tag{3.6}$$

for $a, b \in L^1([0,1])$ upper semicontinuous.

Recall Theorem 5.10 in [23], which states that the following dual Kantorovich formulation equals the Wasserstein distance:

$$W(\mu^u, \mu^0) = \sup_{\substack{(\psi, \psi') \in L_1([0,1])^2 \\ \psi(\gamma_1) - \psi'(\gamma_2) \leq c(\gamma_1, \gamma_2)}} \int_0^1 \psi d\mu^u - \int_0^1 \psi' d\mu^0. \tag{3.7}$$

Define therefore

$$E(u, \psi, \psi') = \int_\Omega f(u(x), x)dx + \lambda \mathrm{TV}(u) + \nu \left(\int_0^1 \psi d\mu^u - \int_0^1 \psi' d\mu^0 \right) \tag{3.8}$$

and let

$$C = \mathrm{BV}(\Omega, [0,1]) \tag{3.9}$$

be the space of functions of bounded variation with domain Ω and range $[0,1]$ and

$$D = \left\{ \psi, \psi' : [0,1] \to \mathbb{R} \text{ s.t. } \begin{array}{c} \psi(\gamma_1) - \psi'(\gamma_2) \leq c(\gamma_1, \gamma_2) \ \forall \gamma_1, \gamma_2 \in [0,1] \\ \psi, \psi' \in L_1([0,1]) \end{array} \right\}. \tag{3.10}$$

It follows from (3.7) with the above definitions that

$$\inf_{u \in C} E(u) = \inf_{u \in C} \sup_{(\psi, \psi') \in D} E(u, \psi, \psi') \tag{3.11}$$

5

3.3 Functional Lifting

While the Wasserstein distance (3.4) is convex in both of its arguments, see Theorem 4.8 in [23], the energy in (3.2) is not convex due to the nonconvex transformation $u \mapsto \mu^u$ in the first argument of the Wasserstein term and the possible nonconvexity of f. To overcome the nonconvexity of both the data term and the transformation in the first argument of the Wasserstein distance we lift the function u. Instead of u we consider a function ϕ defined below whose domain is one dimension larger. This extra dimension represents the range of u and allows us both to linearize the fidelity term and to convexify the Wasserstein distance. This technique, known as functional lifting or the calibration method, was introduced in [1] and is commonly used in many optimization problems.

Let

$$C' = \left\{ \phi \in \mathrm{BV}\left(\Omega \times \mathbb{R}, \{0,1\}\right) \; : \; \begin{array}{c} \phi(\cdot, (-\infty, 0]) \equiv 1, \; \phi(\cdot, [1, \infty)) \equiv 0, \\ D_\gamma \phi(\cdot, \gamma) \leq 0 \end{array} \right\}. \tag{3.12}$$

Every function $u \in C$ corresponds uniquely to a function $\phi \in C'$ via the relation

$$- D_\gamma \phi = \mathcal{H}^2 \llcorner \mathrm{graph}(u), \tag{3.13}$$

where $\mathcal{H}^2 \llcorner \mathrm{graph}(u)$ is the restriction of the 2-dimensional Hausdorff measure to the graph of u. Also for such a pair (u, ϕ) and for all measurable sets $A \subset [0, 1]$ we have the relation

$$\mu^u(A) = \mu^\phi(A) = \frac{1}{|\Omega|} \int_\Omega |D_\gamma \phi(x, A)| dx = \frac{1}{|\Omega|} \int_\Omega -D_\gamma \phi(x, A) dx. \tag{3.14}$$

Note that in contrast to $u \mapsto \mu^u$, the transformation $\phi \mapsto \mu^\phi$ is linear.

Consider the energy

$$E'(\phi, \psi, \psi') = \begin{array}{l} - \int_\Omega \int_0^1 f(\gamma, x) D_\gamma \phi(x, \gamma) \, dx + \lambda \int_0^1 \mathrm{TV}(\phi(\cdot, \gamma)) d\gamma \\ + \nu \left(\int_0^1 \psi d\mu^\phi - \int_0^1 \psi' d\mu^0 \right). \end{array} \tag{3.15}$$

For a pair (u, ϕ) as in (3.13) the identity

$$E(u, \psi, \psi') = E'(\phi, \psi, \psi') \tag{3.16}$$

holds true by the coarea formula, see [2]. Consequently, we have

$$\inf_{u \in C} \sup_{(\psi, \psi') \in D} E(u, \psi, \psi') = \inf_{\phi \in C'} \sup_{(\psi, \psi') \in D} E'(\phi, \psi, \psi'). \tag{3.17}$$

Note that E' is convex in ϕ and concave in (ψ, ψ'), hence is easier to handle from an optimization point of view.

Theorem 1. *Let $\Omega \subset \mathbb{R}^2$ be bounded, let $f(x, \gamma)$ be continuous and let the cost c of the Wasserstein distance fulfill the conditions from Section 3.2. Then there exists a minimizer $\overline{\phi}$ of $\inf_{\phi \in C'} \sup_{(\psi, \psi') \in D} E'(\phi, \psi, \psi')$.*

Proof. We first show that the set C' is compact in the weak* topology in BV. By theorem 3.23 in [2], C' is precompact. It then remains to prove that C' is closed in the weak*-topology. Thus let (ϕ_n) in C' converge weakly* to ϕ, which means that (ϕ_n) converges strongly in L^1_{loc} and $D_\gamma \phi_n$ converges weakly*. $D_\gamma \phi_n(\cdot, \gamma) \leq 0$ means

$$\int_{\Omega \times \mathbb{R}} w D_\gamma \phi_n \geq 0 \quad \forall w \in C_c(\Omega \times \mathbb{R}). \tag{3.18}$$

This property is preserved under weak*-convergence by definition. $\phi(x, \gamma) \in \{0, 1\}$ a.e. as convergence in L^1 implies pointwise convergence of some subsequence. Obviously $\phi_n(\cdot, (\infty, 0]) \equiv 1$ and $\phi_n(\cdot, [1, \infty)) \equiv 0$ are naturally preserved in the limit.

The first term in the energy (3.15) is lower semicontinuous by assumption. The TV-term is lower-semicontinuous by Theorem 5.2 in [2].

The Wasserstein term in (3.15) has the form $\sup_{\{(\psi, \psi') \in D\}} \int_0^1 \psi \, d\mu^\phi - \int_0^1 \psi' \, d\mu^0$ and can thus be written as

$$\sup_{\{(\psi, \psi') \in D, \psi, \psi' \in C_c([0,1])\}} -\frac{1}{|\Omega|} \int_0^1 \int_\Omega \psi(\gamma) dx \, D_\gamma \phi(x, \gamma) - \int_0^1 \psi'(\gamma) \, d\mu^0(\gamma), \tag{3.19}$$

where $C_c([0, 1])$ is the space of all continuous functions in $[0, 1]$. Hence it is a supremum of linear functionals and lsc as well.

As a supremum of positive sums of lsc terms, $\sup_{(\psi, \psi') \in D \cap C_c([0,1])^2} E'(\cdot, \psi, \psi')$ is lsc as well. A minimizing sequence therefore has a weakly*-convergent subsequence due to compactness of C'. The limit is a minimizer by the lower semicontinuity of the energy. $\square$

As we have shown above, the proposed lifted model is well-posed, which means that the minimizer is attained under mild technical conditions. Then by (3.17) also the original energy is well-posed.

Remark 2. We have considered a spatially continuous formulation, as discretizations thereof suffer less from grid bias [10, 11] than purely discrete formulations. Thus, proving existence of a solution of the spatially continuous model substantiates our approach from a modelling point of view.

Remark 3. As discussed in Section 1, we merely consider total variation based regularization as a case study, but this restriction is not necessary. More general regularizers can be used as well as long as they are convex and all the statements still hold, e.g. quadrativ or Huber functions, see [16]. In the present paper however, we rather focus on the novel prior based on the Wasserstein distance.

4 Relaxation as a Convex/Concave Saddle Point Problem

Optimizing energy (3.2) is not tractable, as it is a nonconvex problem. Also solving (3.17) is not tractable, as the set C' is nonconvex. The latter can be

overcome by considering the convex hull of C', which leads to a relaxation as a convex/concave saddle point problem of the minimization problem (3.2), which is solvable computationally.

Proposition 1. *Let*

$$C'' = \{\phi \in BV(\Omega \times \mathbb{R}, [0,1]) \; : \; \phi(\cdot, (-\infty, 0]) \equiv 1, \; \phi(\cdot, [1, \infty)) \equiv 0, \; D_\gamma \phi \leq 0\} \tag{4.1}$$

Then E' is convex/concave and

$$\min_{u \in C} E(u) \geq \min_{\phi \in C''} \sup_{(\psi, \psi') \in D} E'(\phi, \psi, \psi'). \tag{4.2}$$

If

$$\min_{u \in C} \max_{(\psi, \psi') \in D} E(u, \psi, \psi') = \max_{(\psi, \psi') \in D} \min_{u \in C} E(u, \psi, \psi') \tag{4.3}$$

holds, then the above relaxation is exact.

Proof. Note that C'' is a convex set, in particular it is the convex hull of C'. E' is also convex in ϕ, therefore the right side of (4.2) is a convex/concave saddle point problem. For fixed (ψ, ψ') we have the following equality:

$$\min_{u \in C} E(u, \psi, \psi') = \min_{\phi \in C''} E'(\phi, \psi, \psi'), \tag{4.4}$$

which is proved in [15]. Thus

$$
\begin{array}{rcl}
\min_{u \in C} E(u) & = & \min_{u \in C} \sup_{(\psi, \psi') \in D} E(u, \psi, \psi') \\
& \overset{(*)}{\geq} & \sup_{(\psi, \psi') \in D} \min_{u \in C} E(u, \psi, \psi') \\
& \overset{(**)}{=} & \sup_{(\psi, \psi') \in D} \min_{\phi \in C''} E'(\phi, \psi, \psi'),
\end{array}
\tag{4.5}
$$

where $(*)$ is always fulfilled for minimax problems and $(**)$ is a consequence of (4.4). This proves (4.2). If (4.3) holds, then $(*)$ above is actually an equality and the relaxation is exact. $\qquad\square$

5 Relaxation with Hoeffding-Fréchet Bounds

A second relaxation can be constructed by using the primal formulation (3.4) of the Wasserstein distance and enforcing the marginals of the distribution function of the transport plan to be μ^ϕ and μ^0 by the Hoeffding-Fréchet bounds:

Theorem 2 ([18, Thm. 3.1.1])**.** *Let F_1, F_2 be two real distribution functions (d.f.s) and F a d.f. on $\mathbb{R}^2$. Then F has marginals F_1, F_2, if and only if*

$$(F_1(\gamma_1) + F_2(\gamma_2) - 1)_+ \leq F(\gamma_1, \gamma_2) \leq \min\{F_1(\gamma_1), F_2(\gamma_2)\} \tag{5.1}$$

By (3.4) the Wasserstein Distance with marginal d.f.s F_1, F_2 can be computed by solving the optimal transport problem and we arrive at the formulation

$$W(dF_1, dF_2) = \min_F \int_{\mathbb{R}^2} c(dF_1, dF_2) \, dF, \quad \text{s.t. } F \text{ respects the conditions (5.1)}$$

(5.2)

where dF_i shall denote the measure associated to the d.f. F_i, $i = 1, 2$.

Using again the functional lifting technique of [15], the Hoeffding-Fréchet bounds and the representation of the Wasserstein distance (5.2), we arrive at the following relaxation, where we replace the distribution functions F_1 by the distribution function of μ^ϕ, which is $\int_\Omega -D_\gamma \phi(x, [0, \gamma]) dx$.

$$\begin{aligned}
\min_{\phi, F} \quad & \int_\Omega \int_0^1 -f(\gamma, x) D_\gamma \phi(x, \gamma) dx + \lambda \int_0^1 \text{TV}(\phi(\cdot, \gamma) d\gamma + \nu \int_{\mathbb{R}^2} c \, dF, \\
\text{s.t.} \quad & F_\phi(\gamma) = \frac{1}{|\Omega|} \int_\Omega -D_\gamma \phi(x, [0, \gamma]) dx, \\
& F_{\mu^0}(\gamma) = \mu^0([0, \gamma]), \\
& F_\phi(x_1) + F_{\mu^0}(x_2) - 1 \le F(x_1, x_2) \le \min\{F_\phi(x_1), F_{\mu^0}(x_2)\} \\
& \phi \in C''
\end{aligned}$$

(5.3)

The minimization problem (5.3) is a relaxation of (3.2). Just set

$$\phi(x, \gamma) = \begin{cases} 1, & u(x) < \gamma \\ 0, & u(x) \ge \gamma \end{cases}$$

and let F be the d.f. of the optimal transport measure with marginals μ^u and μ^0.

Remark 4. It is interesting to know, when relaxation (5.3) is exact. By the coarea formula [24] we know that

$$\begin{aligned}
& \int_\Omega \int_0^1 -f(\gamma, x) D_\gamma \phi(x, \gamma) dx + \lambda \int_0^1 \text{TV}(\phi(\cdot, \gamma)) d\gamma \\
= \quad & \int_0^1 \int_\Omega f(u_\alpha(x), x) dx d\alpha + \lambda \int_0^1 \text{TV}(u_\alpha) d\alpha,
\end{aligned}$$

(5.4)

where u_α corresponds to the thresholded function $\phi_\alpha = \mathbb{1}_{\{\phi > \alpha\}} \in C'$ via relation (3.13). However such a formula does not generally hold for the optimal transport: Let $\phi_\alpha = \mathbb{1}_{\{\phi > \alpha\}}$ and let F_α be the d.f. of the optimal coupling with marginal d.f.s F_{ϕ_α} and F_{μ^0}. Then

$$F = \int_0^1 F_\alpha \, d\alpha$$

(5.5)

has marginal d.f.s $\int_0^1 F_{\phi_\alpha} d\alpha$ and F_{μ^0}, but it may not be optimal.

A simple example for inexactness can be constructed as follows: Let the data term be $f \equiv 0$ and let $\mu^0 = \frac{1}{2}(\delta_0 + \delta_1)$ and let the cost for the Wasserstein distance be $c(\gamma_1, \gamma_2) = \lambda|\gamma_1 - \gamma_2|$. Every constant function with $u(x) = \text{const} \in [0, 1]$ will be a minimizer if λ is small and ν is big enough. The objective value will be $\frac{\lambda}{2}$. But relaxation (5.3) is inexact in this situation: Choose $\phi(x, \gamma) = \frac{1}{2} \quad \forall \gamma \in (0, 1)$ and the relaxed objective value will be 0.

Remark 5. The above remark was concerned with an example, where a convex combination of optimal solutions to the non-relaxed problem is a unique solution of the relaxed problem with lower objective value.

By contrast, in Section 8 two different academical examples are shown, which illustrate the behaviour of our relaxation (5.3) in situations when the non-relaxed solution is unique, see Figures 2 and 3. Then exactness may hold or not, depending on the geometry of level sets of solutions. No easy characterization seems to be available for the exactness of model (5.3).

6   Relationship between the two Relaxations

Both relaxations from Sections 4 and 5 seem to be plausible but seemingly different relaxations. Their different nature reveals itself also in the conditions for which exactness was established. While the condition in Proposition 1 depends on the gap introduced by interchanging the minimum and maximum operation, relaxation (5.3) is exact if a coarea formula holds for the optimal solution. It turns out, however, that both equations are equivalent, hence both optimality conditions derived in Sections 4 and 5 can be used to ensure exactness of a solution to either one of the relaxed minimization problems.

Theorem 3. *The optimal values of the two relaxations* (4.2) *and* (5.3) *are equal.*

Proof. It is a well known fact that

$$\min_{x \in X} \max_{y \in Y} \langle Kx, y \rangle + G(x) - H^*(y) \qquad (6.1)$$

and

$$\min_{x \in X} H(Kx) + G(x) \qquad (6.2)$$

are equivalent, where $G : X \to [0, \infty]$ and $H^* : Y \to [0, \infty]$ are proper, convex, lsc functions, H^* is the convex conjugate of H and X and Y are two real vector spaces, see [20] for details.

To apply the above result choose

$$G(\phi) = \int_0^1 \int_\Omega -D_\gamma \phi(x, \gamma) \cdot f(\gamma, x) dx + \lambda \int_0^1 \text{TV}(\phi(\cdot, \gamma)) d\gamma + \chi_{C''}(\phi), \quad (6.3)$$

$$H^*(\psi, \psi') = \nu \int_0^1 \psi' d\mu^0 + \chi_D(\psi, \psi') \qquad (6.4)$$

and

$$K : BV(\Omega \times \mathbb{R}, [0, 1]) \to \mathcal{M}([0, 1])^2,$$
$$K(\phi) = (\nu \mu^\phi, 0) \qquad (6.5)$$

where C'' is defined by (4.1), D by (3.10), $\chi_{C''}(\cdot)$ and $\chi_D(\cdot)$ denote the indicator functions of the sets C'' and D respectively and $\mathcal{M}([0, 1])$ denotes the space of

measures on $[0,1]$. (6.1) corresponds with the above choices to the saddle point relaxation (4.2).

Recall that $H = (H^*)^*$ if H is convex and lsc, i.e. H is the Legendre-Fenchel bidual of itself, see [20]. Hence, for positive measures $\mu, \tilde{\mu}$, the following holds true:

$$
\begin{aligned}
H(\mu, \tilde{\mu}) &= \sup_{\psi, \psi'} \{ \int_0^1 \psi d\mu - \int_0^1 \psi' d\tilde{\mu} - H^*(\psi, \psi') \} \\
&= \sup_{(\psi, \psi') \in D} \{ \int_0^1 \psi d\mu - \int_0^1 \psi' d\tilde{\mu} - \nu \int_0^1 \psi' d\mu^0 \} \\
&= \sigma_D(\mu, \tilde{\mu} + \nu \mu^0) \\
&\stackrel{(*)}{=} W(\mu, \tilde{\mu} + \nu \mu^0)
\end{aligned}
\tag{6.6}
$$

where $\sigma_A(x) = \sup_{a \in A} \langle a, x \rangle$ is the support function of the set A and ν is the weight for the Wasserstein term in (3.8). To prove $(*)$, we invoke Theorem 5.10 in [23], which states that

$$
\sigma_D(\mu, \tilde{\mu}) = \sup_{(\psi, \psi') \in D} \int_0^1 \psi d\mu - \int_0^1 \psi' \tilde{\mu} = \min_{\pi \in \Pi(\mu, \tilde{\mu})} \int_{[0,1]^2} c(\gamma_1, \gamma_2) d\pi(\gamma_1, \gamma_2) = W(\mu, \tilde{\mu}),
\tag{6.7}
$$

and we have infinity for measures which do not have the same mass.

Thus, the energy in (6.2) can be written as

$$
G(\phi) + H(\nu \mu^\phi, 0) = G(\phi) + W(\nu \mu^\phi, \nu \mu^0) = G(\phi) + \nu W(\mu^\phi, \mu^0).
\tag{6.8}
$$

This energy is the same as in relaxation (5.3), which concludes the proof. $\qquad\square$

7 Optimization

We present five experiments and the numerical method used to compute them.

7.1 Implementation

First, we discretize the image domain Ω to be $\{1, \ldots, n_1\} \times \{1, \ldots, n_2\}$ and use forward differences as the gradient operator. Second, we discretize the infinite dimensional set C'' and denote it by

$$
C''_d = \left\{ \phi : \Omega \times \left\{ 0, \frac{1}{k}, \ldots, \frac{k-1}{k}, 1 \right\} \to [0,1] : \begin{array}{l} \phi(\cdot, 1) = 0, \ \phi(\cdot, 0) = 1, \\ \phi\left(\cdot, \frac{l}{k}\right) \le \phi\left(\cdot, \frac{l-1}{k}\right) \end{array} \right\}.
\tag{7.1}
$$

Hence we consider only finitely many grayvalues an image can take. The dual Kantorovich set for the discretised problem is then

$$
D_d = \left\{ \psi, \psi' : \left\{ 0, \frac{1}{k}, \ldots, \frac{k-1}{k}, 1 \right\} \to \mathbb{R} : \quad \psi(\gamma_1) - \psi'(\gamma_2) \le c(\gamma_1, \gamma_2) \ \forall \gamma_1, \gamma_2 \right\}.
\tag{7.2}
$$

After computing a minimizer ϕ^* of the discretized energy, we threshold it at the value 0.5 to obtain $\phi^* = \mathbb{1}_{\{\phi^* > 0.5\}}$ and then calculate $u^\star$ by the discrete analogue of relation (3.13).

11

For computing a minimizer of the discretized optimization problem

$$\min_{\phi \in C''_d} \max_{(\psi,\psi') \in D_d} E'_d(\phi, \psi, \psi') \tag{7.3}$$

it is expedient to use first order algorithms like [3,5,7,19] as the dimensionality of the problem is high. To use such algorithms it is necessary to split the function $\max_{(\psi,\psi') \in D_d} E'_d(\phi, \psi, \psi')$ into a sum of terms, whose proximity operators can be computed efficiently. Hence consider the following equivalent minimization problem:

$$\min_{\phi \in C''_d, g \in (\mathbb{R}^{n1 \times n2 \times k \times 2})} \langle \tilde{f}, \phi \rangle + \|g\|_1 + \chi_{\{(u,v):\, \nabla u = v\}}(\phi, g) + \chi_{C''}(\phi) + W(\mu^\phi, \mu^0),$$
$$\tag{7.4}$$

where $\tilde{f}$ comes from the local cost factor in (3.15) and χ_A is the indicator function of a set A. The proximity operator of a function G is defined as

$$\text{prox}_G(x) = \text{argmin}_{x'} \frac{1}{2}\|x - x'\|^2 + G(x'). \tag{7.5}$$

The proximity operator of the term $\|g\|_1$ is the soft-thresholding operator.

$\text{prox}_{\chi_{\{(u,v):\, \nabla u = v\}}}(\phi, g)$ can be efficiently computed with Fourier transforms, see for example [19].

$\text{prox}_{\chi_{C''}}$ is the projection onto the set of non-increasing sequences C''. To compute this projection, we employ the algorithm proposed in [4], Appendix D. It is trivially parallelisable and converges in a finite number of iterations.

Finally, the proximity operator for the Wasserstein distance can be computed efficiently in some special cases, as discussed in the next Section 7.2.

We can either use [19] to minimize (7.5) directly, which is equivalent to using the Douglas-Rachford method [7] on a suitably defined product space and absorbing the linear term in the functions in (7.5).

7.2 Wasserstein Proximation for $c(\gamma_1, \gamma_2) = |\gamma_1 - \gamma_2|$ by soft-thresholding

In general, computing the proximity operator for the Wasserstein distance can be expensive and requires solving a quadratic program. However, for the real line and convex costs, we can compute the proximity operator more efficiently. One algorithm for the cost function $c(\gamma_1, \gamma_2) = |\gamma_1 - \gamma_2|$ is presented below.

The proximation for the weighted Wasserstein distance is

$$\text{argmin}_\phi \frac{1}{2}\|\phi^0 - \phi\|_2^2 + \lambda W(\mu^\phi, \mu^0). \tag{7.6}$$

For the special case we consider here, there is a simple expression for the Wasserstein distance:

Proposition 2 ([18]). *For two measures μ^1, μ^2 on the real line and $c(\gamma_1, \gamma_2) = |\gamma_1 - \gamma_2|$, the Wasserstein distance is*

$$W(\mu^1, \mu^2) = \int_{\mathbb{R}} |F_{\mu^1}(\gamma) - F_{\mu^2}(\gamma)| \, d\gamma \tag{7.7}$$

12

Due to $D_\gamma\phi(x,\gamma) \le 0$ and $\phi(x,0) = 1$, we can also write $F_{\mu^\phi}(\gamma)$ as

$$F_{\mu^\phi}(\gamma) = \frac{1}{|\Omega|}\int_\Omega -D_\gamma\phi(x,[0,\gamma])\ dx = \frac{1}{|\Omega|}\int_\Omega 1 - \phi(x,\gamma)dx\,. \tag{7.8}$$

Next we show how to solve in closed form the proximity operator for the Wasserstein distance in the present case.

Proposition 3. *Given ϕ^0, $\lambda > 0$, the optimal $\tilde\phi$ for the proximity operator*

$$\tilde\phi = \mathrm{argmin}_\phi\ \frac{1}{2}\|\phi - \phi^0\|_2^2 + \lambda W(F_{\mu^\phi},\mu^0) \tag{7.9}$$

is determined by

$$\tilde\phi(x,\gamma) = \phi(x,\gamma) + c_\gamma\,, \tag{7.10}$$

where

$$c_\gamma = shrink\left(-\frac{1}{|\Omega|}\int_\Omega \phi^0(x,\gamma)dx - F_{\mu^0}(\gamma) + 1, \frac{\lambda}{|\Omega|}\right) + \frac{1}{|\Omega|}\int_\Omega \phi^0(x,\gamma)dx + F_{\mu^0}(\gamma) - 1 \tag{7.11}$$

and shrink denotes the soft-thresholding operator defined componentwise by

$$shrink(a,\lambda)^i = (|a^i| - \lambda)_+ \cdot \mathrm{sign}(a^i) \tag{7.12}$$

for $a \in \mathbb{R}^n$, $\lambda > 0$.

Proof. By proposition 2 and the characterisation of F_{μ^ϕ} in (7.8), proximation (7.6) reads

$$\mathrm{argmin}_\phi\ \frac{1}{2}\|\phi^0 - \phi\|_2^2 + \lambda\int_\mathbb{R} |1 - \left(\frac{1}{|\Omega|}\int_\Omega \phi(x,\gamma)dx\right) - F_{\mu^0}(\gamma)|d\gamma. \tag{7.13}$$

Note that (7.13) is an independent optimization problem for each γ. Thus, for each γ we have to solve the problem

$$\mathrm{argmin}_{\phi(\cdot,\gamma)}\ \frac{1}{2}\|\phi^0(\cdot,\gamma) - \phi(\cdot,\gamma)\|_2^2 + \lambda|1 - \left(\frac{1}{|\Omega|}\int_\Omega \phi(x,\gamma)dx\right) - F_{\mu^0}(\gamma)|. \tag{7.14}$$

It can be easily verified that the solution to problem (7.14) is $\phi^0(\cdot,\gamma) + c_\gamma$, where $c_\gamma \in \mathbb{R}$ and

$$c_\gamma \in \mathrm{argmin}_{c\in\mathbb{R}}\ \frac{1}{2}|\Omega|c^2 + \lambda|\frac{1}{|\Omega|}\int_\Omega \phi^0(x,\gamma)dx + c + F_{\mu^0}(\gamma) - 1| \tag{7.15}$$

and hence

$$c_\gamma = shrink\left(-\frac{1}{|\Omega|}\int_\Omega \phi^0(x,\gamma) - F_{\mu^0}(\gamma) + 1, \frac{\lambda}{|\Omega|}\right) + \frac{1}{|\Omega|}\int_\Omega \phi^0(x,\gamma)dx + F_{\mu^0}(\gamma) - 1. \tag{7.16}$$

$\square$

For the discretized problem one just needs to replace integration with summation to obtain the proximation operator. Concluding, the cost for the Wasserstein proximal step is linear in the size of the input data.

Remark 6. We have seen in Proposition 2 that $W_1(\mu^\phi, \mu^0) = \|H\phi - (1 - F\mu^0)\|_1$, where H is an operator corresponding to a tight frame, i.e. $HH^* = |\Omega|^{-1}$, hence it is also possible to derive Proposition 3 by known rules for proximity operators involving composition with tight frames and translation.

8   Numerical Experiments

We want to show experimentally

1. that computational results conform to the mathematical model,

2. that the convex relaxation is reasonable.

Note that we do not claim to achieve the best denoising or inpainting results and we do not wish to compete with other state-of-the-art methods here. We point out again that the Wasserstein distance can be used together with other variational approaches to enhance their performance, e.g. with nonlocal total variation based denoising, see [9].

Remark 7. As detailed in Section 3.3, we lift our functional, so that it has one additional dimension, thereby increasing memory requirements and runtime of our algorithm. Non-convex approaches like [17] do not have such computational requirements. Still, the viability of the lifting approach we use was demonstrated in [16] for our variational model without the Wasserstein term. Also all additional operations our algorithm requires can be done very fast on recent graphic cards, hence the computational burden is tractable.

We have generally chosen the parameters λ, ν by hand to obtain reasonable results, if not stated differently.

In the **first experiment** we compare total variation denoising and total variation denoising with the Wasserstein term for incorporating prior knowledge. The data term is $f(s, x) = (u_0(x) - s)^2$, where u_0 is the noisy image in figure 1. The cost for the Wasserstein distance is $c(\gamma_1, \gamma_2) = \nu|\gamma_1 - \gamma_2|$, $\nu > 0$. To ensure a fair comparison, the parameter λ for total variation regularization *without* the Wasserstein term was *hand-tuned in all experiments* to obtain best results. The histogram was chosen to match the noiseless image. See Figure 1 for the results.

Note the trade-off one always has to make for pure total variation denoising: If one sets the regularization parameter λ high, the resulting grayvalue histogram of the recovered image will be similar to the noisy input image and generally far away from the histogram of ground truth. By choosing lower data fidelity and higher regularization strength we may obtain a valid geometry of the image, however then the grayvalue histogram tends to be peaked at one mode, as total variation penalizes scattered histograms and tries to draw the
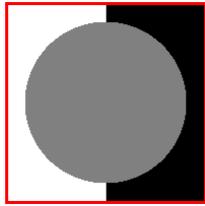
(a) The gray area is to be inpainted with partly black and white, with slightly more white.

(b) The circle in the middle has been inpainted with slightly more white as demanded by the Wasserstein term.
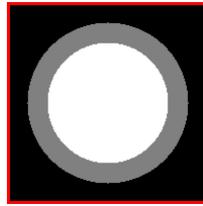
(c) The gray area is the area to be inpainted with a given Wasserstein prior favoring the gray area to be half black and half white.
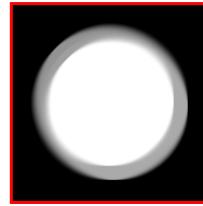
(d) Inpainting result: we obtain a non-integral solution visualized by gray color.

Figure 2: Example illustrating tightness of our relaxation (5.3).

Figure 3: Example illustrating failure of tightness of our relaxation (5.3).

modes closer to each other, again letting the recovered grayvalue histogram being different from the desired one. By contrast, the Wasserstein prior in (3.8) guarantees a correct grayvalue histogram also with strong spatial regularization.

The **second** set of experiments illustrates where exactness of our relaxation may hold or fail, depending on the geometry of the level sets of solutions, see Figures 2 and 3. The gray area is to be inpainted with a Wasserstein prior favoring the gray area to be partly black and partly white. Note that both settings illustrate cases, when the global Wasserstein term is indispensable, as otherwise there would be completely no control over how much of the area to be inpainted ends up being white or black. While our relaxation is not exact for the experiment in Figure 3, thresholding at 0.5 still gives a reasonable result.

The **third** experiment is a more serious denoising experiment. Notice that again pure total variation denoising does not preserve the white and black areas well, but makes them gray, while the approach with the Wasserstein distance preserves the contrast better, see Figure 4.

In the **fourth experiment** we compare image inpainting with a total variation regularization term without prior knowledge and with prior knowledge, see Figure 5 for the results. The region where the data term is zero is enclosed in the blue rectangle. Outside the blue rectangle we employ a quadratic data term as in the first experiment. Total variation inpainting without the Wasserstein term does not produce the results we expected, as the total variation term is smallest, when the gray color fills most of the area enclosed by the blue rectangle. Heuristically, this is so because the total variation term weighs the boundary length multiplied by the difference between the gray value intensities, and a medium intensity minimizes this cost. Thus the TV-term tends to avoid interfaces, where high and low intensities meet, preferring smaller intensity changes, which can be achieved by interfaces with gray color on one side. Note that also the regularized image with the Wasserstein term lacks symmetry. This is also
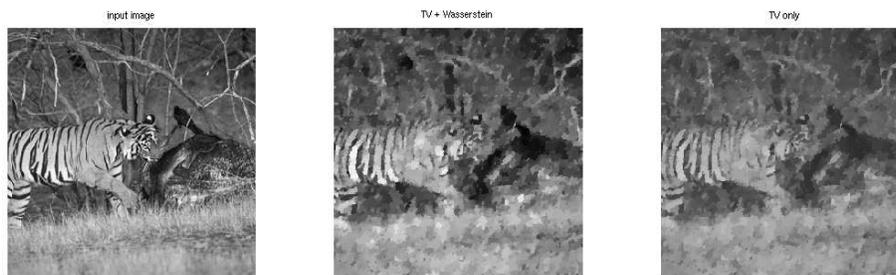
(a) Tiger denoising experiment with the original image on the left, the image denoised with the Wasserstein term in the middle and the standard ROF-model on the right.
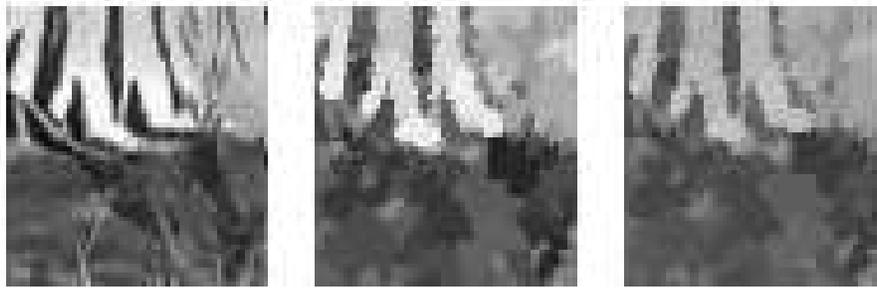
(b) Detailed view of the tiger denoising experiment revealing that contrast is better preserved when the Wasserstein term is used.

Figure 4: Tiger denoising experiment
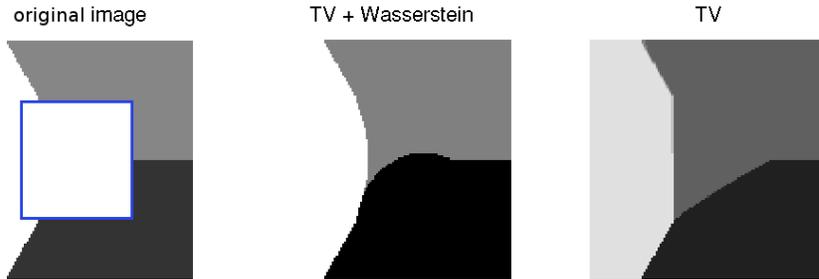
Figure 5: Inpainting experiment with the original image and the inpainting area enclosed in a blue rectangle on the left, the inpainting result with the Wasserstein term in the middle and the result where only the TV-regularizer is used on the right. By enforcing the three regions to have the same size with the Wasserstein term, we obtain a better result than with the Total Variation term alone.
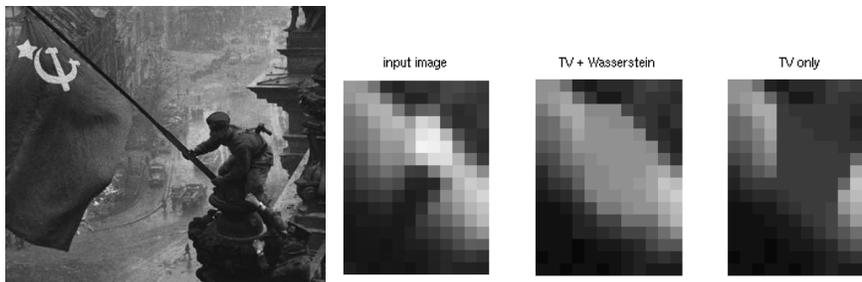

Figure 6: Here we want to inpaint the area occupied by the watch of the soldier, see the second left image. Our approach, on the second right image gives better results again than the approach with TV alone.

due to the behaviour of the TV-term described above.

In the **fifth** experiment we consider inpainting again. Yevgeni Khaldei, the photographer of the iconic picture shown on the left of Figure 6 had to remove the second watch. Trying to inpaint the wrist with a TV-regularizer and a Wasserstein term results in the middle picture, while only using a TV-regularizer results in the right picture. Clearly using the Wasserstein term helps.

In the **sixth** experiment we have a different setup. The original image is on the left of Figure 7. The histogram μ^0 was computed from a patch of clouds, which did not include the plane. The data term is $f(x,y) = \lambda \min(|u_0(x) - y|^2, \alpha)$, where $\alpha > 0$ is a threshold, so the data term does not penalize great deviances from the input image too strongly. The Wasserstein term penalizes the image of the plane whose appearance differs from the prior statistics. The TV-regularizer is weighted weaker than in the previous examples, because we do not want to smooth the clouds.

Figure 7: *Unsupervised* inpainting using empirical measures as priors. Objects not conforming to the prior statistics are removed *without* labeling image regions.

Note that unlike in ordinary inpainting applications, we did not specify the location of the plane beforehand, but the algorithm figured it out on its own. The total variation term finally favors a smooth inpainting of the area occupied by the plane. In essence we have combined two different tasks: Finding out where the plane is and inpainting that area occupied by it. See Figure 7 for results.

9   Conclusion and Outlook

We have presented in this paper a novel method for variational image regularization, which takes into account global statistical information in one model. By solving the relaxed nonconvex problem we obtain regularizd images which conform to some global image statistics, which sets our method apart from standard variational methods. Moreover, the additional computational cost for the Wasserstein term we introduced is negligible, however our relaxation is not tight anymore as in models without the latter term. In our experiments the relaxation was seen to be tight enough for good results.

Our future work will consider extensions of the present approach to multidimensional input data and related histograms, e.g. based on color, patches or gradient fields. The theory developed in this paper regarding the possible exactness of solutions does not carry over without modifications to such more complex settings. Moreover, it is equally important to find ways related to our present work to minimize such models efficiently.

Acknowledgements. We would like to thank the anonymous reviewers for their constructive criticism and Marco Esquinazi for helpful discussions.

References

[1] G. Alberti, G. Bouchitte, and G. Dal Maso. The calibration method for the Mumford-Shah functional and free-discontinuity problems. *Journal: Calc. Var. Partial Differential Equations*, 16:299–333, 2003.

[2] L. Ambrosio, N. Fusco, and D. Pallara. *Functions of Bounded Variation and Free Discontinuity Problems (Oxford Mathematical Monographs)*. Oxford University Press, USA, May 2000.

[3] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Found. Trends Mach. Learning*, 3(1):1–122, 2010.

[4] A. Chambolle, D. Cremers, and T. Pock. A Convex Approach to Minimal Partitions. *SIAM J. Imag. Sci.*, 5(4):1113–1158, 2012.

[5] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

[6] T. F. Chan, S. Esedoglu, and K. Ni. Histogram based segmentation using Wasserstein distances. In Fiorella Sgallari, Almerico Murli, and Nikos Paragios, editors, *SSVM*, volume 4485 of *Lecture Notes in Computer Science*, pages 697–708. Springer, 2007.

[7] J. Eckstein and D. P. Bertsekas. On the Douglas—Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55:293–318, 1992.

[8] S. Ferradans, G-S. Xia, G. Peyré, and J-F. Aujol. Optimal transport mixing of gaussian texture models. In *Proc. SSVM'13*, 2013.

[9] G. Gilboa and S. Osher. Nonlocal Operators with Applications to Image Processing. *Multiscale Modeling & Simulation*, 7(3):1005–1028, 2008.

[10] M. Klodt, T. Schoenemann, K. Kolev, M. Schikora, and D. Cremers. An experimental comparison of discrete and continuous shape optimization methods. In *Proceedings of the 10th European Conference on Computer Vision: Part I*, ECCV '08, pages 332–345, Berlin, Heidelberg, 2008. Springer-Verlag.

[11] J. Lellmann, B. Lellmann, F. Widmann, and C. Schnörr. Discrete and Continuous Models for Partitioning Problems. *Int. J. Comp. Vision*, 2013. in press.

[12] J. Lellmann and C. Schnörr. Continuous Multiclass Labeling Approaches and Algorithms. *SIAM J. Imag. Sci.*, 4(4):1049–1096, 2011.

[13] N. Paragios, Y. Chen, and O. Faugeras, editors. *The Handbook of Mathematical Models in Computer Vision*. Springer, 2006.

[14] G. Peyré, J. Fadili, and J. Rabin. Wasserstein active contours. In *Proc. ICIP'12*, 2012.

[15] T. Pock, D. Cremers, H. Bischof, and A. Chambolle. Global Solutions of Variational Models with Convex Regularization. *SIAM J. Imaging Sciences*, 3(4):1122–1145, 2010.

[16] T. Pock, T. Schoenemann, G. Graber, H.Bischof, and D. Cremers. A Convex Formulation of Continuous Multi-label Problems. In *ECCV (3)*, pages 792–805, 2008.

[17] J. Rabin and G. Peyré. Wasserstein regularization of imaging problem. In *ICIP*, pages 1541–1544, 2011.

[18] S. T. Rachev and L. Rüschendorf. *Mass Transportation Problems. Vol. I, Theory*. Springer-Verlag, New York, 1998.

[19] H. Raguet, J. Fadili, and G. Peyré. Generalized forward-backward splitting. Technical report, Preprint Hal-00613637, 2011.

[20] R. T. Rockafellar. *Convex Analysis*. Princeton Landmarks in Mathematics,. Princeton University Press, Princeton, princeton paperbacks edition, 1997.

[21] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *Int. J. Comput. Vision*, 40(2):99–121, November 2000.

[22] L. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1-4):259–268, 1992.

[23] C. Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer, 1 edition, November 2008.

[24] W.P. Ziemer. *Weakly Differentiable Functions*. Springer, 1989.